

Resolutions to enhance confident creativity

Outstanding discoveries are inspiring, thanks to the research opportunities that they open up. But the creative processes involved, and obstacles to them, also deserve consideration.

The laser and its microwave predecessor, the maser, shouldn't have worked. No less a giant than Niels Bohr (and many other physicists) said so — a beam with such a narrow frequency bandwidth would contravene the uncertainty principle. No less an expert than I. I. Rabi (and many others) said so: the second law of thermodynamics would require the generating molecules to be so hot that they would dissociate. Rabi, then the head of the physics department at Columbia University, tried vigorously to persuade the discoverer of the principle on which masers and lasers are based, a member of his department, to cease work on the subject. As Charles Townes subsequently wrote (*How the Laser Happened*; Oxford University Press, 1999), it was a good thing he had tenure, as well as the licence from his funding sources to pursue his ideas for their own sake rather than for specified outcomes.

Necessity was the mother of Townes' invention: a committee he chaired had failed to make much progress in identifying new ways of generating millimetre radiation, and the need to do better led him to go back to first principles. Despite the subsequent opposition, he went on to validate his convictions that neither the uncertainty principle nor the second law were obstacles to unprecedentedly powerful and finely tuned radiation from a stimulated population of molecules; he constructed the first maser in 1954. The Russian theorist Vitali Ginzberg wondered why it hadn't been invented decades earlier, when all the necessary fundamental physics was already established.

The sort of opposition that Townes encountered is inevitable — there will always be established and influential scientists who resent the incursions of new thinkers on their turf or, even with the best will in the world, are resistant to looking beyond their long-held scientific assumptions. And anyway, critical scrutiny is part of the bedrock of science. Regrettably, not everyone is given a Nobel prize-winning combination of experience, opportunity, insight and persistence in the face of such resistance to originality. Nevertheless, it is worth asking whether one can draw inspiration from such examples in trying to ensure that scientifically creative thinking is actively developed in the young, supported rather than opposed at maturity, and sustained rather than undermined at all stages of a career.

Understanding creativity

Psychologists have studied creativity systematically since the 1950s. Some concede that, for all its diversity of approaches, the discipline has delivered few solid new insights (for a comprehensive round-up, see *Handbook of Creativity*, ed. R. Sternberg; Cambridge University Press, 1999). Still, Townes' undramatic but compelling account of his varied and war-stimulated career before his discovery and the work he had to do to fulfil it provides an excellent example of the psychologists' standard model of the creative process: a period of (often unwitting) preparation, in which the creator becomes intimately familiar with an intellectual or practical domain; incubation, in which knowledge and ideas simmer, often while the creator's mind is directed elsewhere;

inspiration — the forever mysterious moment in which the concept of a creation is sparked; and, finally, implementation — the phase in which, in many cases, scepticism needs to be overcome, persuasiveness in gathering support is vital, and, in all cases, time needs to be available for intense, unremitting effort.

Seeking problems

Despite the lack of progress in understanding creativity, there is no reason to doubt that it can be fostered by expecting people to seek problems as well as to solve them, and by making imagination, rather than the products of it, a primary goal in part of their training. Little can be done about inspiration, but there is a worthwhile challenge, rarely confronted in the education and training of scientists, in helping with other aspects of the process for those many individuals who have the capacity to think of productive new associations of ideas, or better ways of doing things, or even fundamentally new concepts. So here are three idealistic resolutions for the year 2000.

- Develop an undergraduate option designed to encourage not only the creative answering of questions but also the creative finding of them in the first place. Take a lead from the "General paper" — the most challenging exam for students of some physics courses: it demands solutions to such unexpected challenges as "derive a theory for the fact that an inverted soy-sauce bottle delivers a small quantity of sauce and then stops, and predict the quantity delivered". (For this, and other general problems, see *Thinking Like a Physicist*, ed. N. Thompson; Hilger, 1987.) Tackling equivalent teasers in any discipline will at least help develop flexible thinking.

- But go further by encouraging students to look for problems. For example, invite students to find everyday phenomena that raise interesting but soluble conundrums. If that opens up avenues for investigation, give the student (whose self-motivation is all-important) a chance to become familiar with the relevant science and its literature, to incubate thoughts, to have ideas and to pursue them. In responding to those ideas, highlight their positive aspects, but also familiarize their originators with the challenge of constructive but tough questioning.

- Setting people goals other than the exploration of their own ideas for its own sake limits creativity. Targets of publication output and other productivity measures undermine creativity psychologically as well as inducing a short-term, problem-solving outlook. Strive to reduce rather than increase the burden of such millstones.

- Reflect on the fact that if Charles Townes had not had tenure, the development of the laser might well have been delayed yet further. That too many of today's creative scientists lack long-term security is a well-known problem; the corollary is that good but unconventional ideas are probably falling by the wayside. There has been much agonizing over, but little progress in, improving the status of contract researchers. Anyone in a luckier position of influence should take at least one initiative this year to help improve their financial independence or otherwise boost their freedom of creative action.

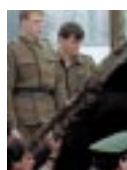




Idaho
Pork-barrel politics
actually works in
funding US science
p5



BSE
Why the public's
respect for experts
went up in flames
p6



Berlin
Winners and losers
in the changing post-
Communist world
p8



Clone
If science is king,
should we temper
its authority?
p9

Europe lifts patent embargo on transgenic plants and animals...

Munich

The European Patent Office (EPO) has lifted a four-year moratorium on applications for patents on plants and animals, following a ruling by its top appeals board that such patents are not excluded by the wording of the European Patent Convention.

In 1995, a lower board of appeal overruled a patent issued to the Belgian company Plant Genetics Systems on a process for creating plants containing herbicide-resistance genes (see *Nature* **374**, 8; 1995). The board's key reason was that the procedure could be used to create new 'varieties' of plants and that plant varieties were excluded from patentability by a clause in the convention.

A similar reason was used by the EPO — which is not part of the European Union but is responsible for implementing the patent convention — to reject an application by the Swiss company Novartis to patent a process for creating transgenic plants containing pathogen-resistance genes, an application which included the plants produced by this process. The EPO president eventually referred Novartis's appeal to the organization's highest legal authority, the 'enlarged board of appeal', and asked for its interpretation of the crucial clause 53(b).

In a ruling made public on 20 December, the enlarged board overturned the earlier judgement. It ruled that the exclusion of plant varieties should be interpreted in a narrow sense, not used to deny patents on all plants produced by novel — and hence patentable — biotechnological processes.

Although this ruling brings the implementation of the patent convention in line with recent modifications to the European Union directive on patenting biotechnological processes (see *Nature* **388**, 314 & **393**, 200; 1998), the enlarged board made little reference to the directive. It argued from first principles that the exclusion of plant varieties from patent protection had a specific function: to let breeders continue applying their own procedures for plant variety protection, as defined by the International Union for the Protection of New Varieties of Plants.

The board considered that the wording of the patent convention, approved in the early



1970s, had been drawn up deliberately to ensure that plants could be protected either through patents or through traditional plant breeders' rights. It therefore concluded that there had been no intention to exclude plants (or animals) from patent protection as such.

The EPO has already started to process

some 1,200 applications for patents on genetically altered animals and plants that have been on hold since 1995.

"This is a landmark decision for the whole seeds industry," says Walter Smolders, head of the patenting department of Novartis Seeds. He says that the "end to uncertainty about intellectual property" will promote development of useful crops and vegetables.

But the decision has been criticized by the environmental group Greenpeace, which campaigns against the patenting of life forms as "presumptuous and reprehensible".

Stefan Flothmann, head of the genetic engineering department at Greenpeace Germany, warns that the decision could lead to monopolies in the seeds market, and thus to higher costs and lower quality.

"If the production of seeds and food gets under control by few companies, the agricultural diversity will be reduced to a few patented varieties," says Flothmann. "In the long term, this is a threat to the world's food supplies." **Quirin Schiermeier and David Dickson**

... as US tightens up on 'speculative' claims

Washington

The US Patent and Trademark Office has released revised guidelines for patent inspectors that require applicants to "explicitly identify a specific and substantial utility" for genes and gene products — in other words, to specify an immediate and identifiable benefit to the public.

The guidelines were revised because of concerns that "people would claim utilities for which they had no substantial showing", says Todd Dickinson, the office's commissioner on patents and trademarks. Past applicants sometimes sought claims on 'naked' expressed sequence tags, he says, for which no known function was attributed. Those

claims were rejected as too speculative, says Dickinson.

Some claimed 'throw-away' utility, saying that a protein could potentially be used as a food supplement or a shampoo additive. Applications quoting an unproven utility will now be rejected.

"We're trying to strike the right balance between the protection of innovation that these inventions represent and the access that the world — and in particular the research community — needs," says Dickinson.

Patent examiners are applying the new guidelines to applications already in the pipeline, including some that were filed provisionally.

Celera Genomics of Rockville, Maryland, announced last October

that it had filed 6,500 provisional applications on genes it discovered while sequencing its first billion base pairs of human DNA. The subset it sends for examination will be subject to the more stringent rules. Paul Gilman, Celera's director of policy and planning, says they "both expect and welcome" the new scrutiny.

Harold Varmus, outgoing director of the US National Institutes of Health, applauds the change. "I've been very pleased with the way [the patent office has] come closer to our position about the need to define specific utility," he says. The guidelines, available at <http://www.uspto.gov>, are open for public comment until 22 March.

Paul Smaglik

Big gains for biology and IT in Japanese science budget

Tokyo

Japan will spend a record amount on science and technology in the next fiscal year, thanks to the nation's largest-ever budget. The budget, approved by the Cabinet on 24 December, includes big increases in support for the life sciences and information technology.

In the financial year 2000, beginning on 1 April, government ministries will have ¥1.2 trillion (US\$11.8 billion) to spend on research, a rise of 6.8 per cent.

The increase is largely the result of funding for prime minister Keizo Obuchi's Millennium Project, launched last year to promote industries in areas such as biotechnology, information sciences and the environment (see *Nature* **401**, 3; 1999).

These projects account for roughly 10 per cent of the new spending, of which ¥64 billion is allocated to research on human and rice genomes. The rest will be spent on the development of high-speed Internet services and environmental technology.

The total government budget will grow by 3.8 per cent to ¥84.9 trillion. Despite a pledge to target areas such as economic structural reform and environmental planning, the budget is still dominated by large public-works projects.

The budget has been criticized for its heavy reliance on government bonds; the national debt is expected to grow to ¥645 trillion by next year, 1.3 times the country's gross domestic product.

Although many have welcomed the support for science and technology, an official from the Science and Technology Agency (STA) says that the budget "is bad news for tax payers", as the overall increase in the agency's budget, at 0.8 per cent, is relatively modest compared to past increases.

Spending on genome research, however, will grow by almost 50 per cent, to ¥17.6 billion. Of this, ¥10 billion will go to the Genome Sciences Centre at the Institute of Physical and Chemical Research (RIKEN), and ¥1.9 billion towards setting up a centre for medical research on single-nucleotide polymorphisms.

A planned national centre for regenerative medicine, which would do research on the development, differentiation and regeneration of plant and animal cells, will also get ¥61 billion for building its facilities.

In contrast, the budgets for space development and nuclear research will be cut by 8.9 per cent and 3.9 per cent, respectively. Last year's failed launch of the H-II rocket (see *Nature* **402**, 336; 1999) caused the government to abandon the H-II programme

Japan's science and technology budget fiscal year 2000 (in billion yen; US\$1 = ¥103)

Science and Technology Agency		
	Budget	Change from 1999
Total budget	780.3	+ 0.8%
Genome research	17.6	+ 47.9%
Brain research	20.1	+ 9.8%
Space development	170.6	- 8.9%
Nuclear energy (including nuclear safety)	257.4	- 3.9%

Ministry of International Trade and Industry

Total R&D budget	528.3	+ 4.4%
Biotechnology	10.3	New
Information technology	10.4	New
Industry/university collaboration	16.9	+ 100%
Nuclear safety	127.8	+ 22.9%

Ministry of Education, Science, Sports and Culture

Grants-in-aid for scientific research	141.9	+ 8.0%
Joint research with industry	120.9	+ 2.7%
Postdoctoral fellowships	17.9	+ 1.9%
Promotion of biological sciences	42.9	+ 56.5%
Centre of excellence programme	70.2	- 0.5%

and focus on the H-IIA rocket, for which it has allocated ¥52 billion in the 2000 budget.

Support for start-up companies and businesses spun-off from research, particularly from universities, also features strongly in the budget. The STA, the Ministry of Education, Science, Sports and Culture (Monbusho) and the Ministry of International Trade and Industry (MITI) will all increase support for collaborations between industry, government research institutes and universities.

Monbusho's postdoctoral fellowship scheme, which reached its target of increasing the ranks of Japan's postdoctoral researchers to 10,000 by this year, will support an additional 400 postdocs with the STA.

The fellowship scheme has been criticized for increasing the number of postdoctoral researchers without creating the research posts to accommodate them. But MITI hopes to promote collaboration between young researchers and venture businesses.

MITI's budget includes money for the analysis of human and microbial genomes, bioinformatics and the analysis of protein structure and function.

The ministry will also receive money for nuclear safety, including the regulation of high-level nuclear waste.

Asako Saegusa

Genomics companies boom on New York stock exchange

Washington

A clutch of US biotech companies with genomics-based business plans saw their stock prices rise dramatically in heavy trading in the final weeks of December. Some reached almost ten times their initial value.

According to biotech analyst Steven Burrill of San Francisco, the rise is the result of a "confluence" of factors — including the completion of human chromosome 22, the sequencing of the first billion base pairs of the human genome and the impending completion of the *Drosophila* genome.

Celera Genomics Inc., which is competing with the international Human Genome Project to sequence the human genome, rose to \$186 a share by the beginning of this week, more than three times its price at the beginning of December.

Paul Boni, an analyst with Punk, Ziegel and Co. in New York, suggests that the sequencing race may have helped to raise both Celera's profile and its price. "Who wins this race is largely irrelevant," he says, adding that it has boosted the fortunes of other genomics database companies. Incyte Pharmaceuticals Inc., of Palo Alto, California, has seen its price climb from \$12 a share in November to \$89 by 3 January.

David Gardner, co-founder of 'The Motley Fool', an online investment-strategies site, says that the addition of Celera's stock to their portfolio — the site purchased the stock at \$79 in mid-December — was "an obvious catalyst" in its rise.

Eric Schmidt, a biotech analyst with SG Cowen Securities in New York, says that Pfizer's decision in November to subscribe to genome databases owned by Celera and Incyte could also have triggered the increase.

Schmidt points out that Celera and Incyte differ from traditional biotech companies in that their futures do not depend on getting a compound through clinical trials, but on supplying the "picks and shovels" for gene prospecting.

Several companies that use genomics to develop potential tests or therapies have also done well. These include Human Genome Sciences Inc. of Rockville, Maryland, whose shares rose from around \$40 in August to almost \$145 last week. For long-term success, however, "they will have to come up with a blockbuster", says Schmidt. Paul Smaglik

From our own correspondents

The beginning of a new year, a new decade — or even a new millennium — is a time for analysis and prognosis. On the next five pages, *Nature's* senior news correspondents reflect on their experiences in reporting from different parts of the world during the past decade. To these recollections, they add thoughts about the current state of play between science and society across the world, and brief personal reflections on how this relationship is likely to develop in the years ahead.

'US politicians have fallen in love with basic research'



Colin Macilwain is the senior US correspondent for *Nature*, and has been based in Washington since 1993.

Washington

A picture of America that sticks in my mind, after seven years of covering US science policy from Washington, is of a plateau in southern Idaho. I was approaching the Argonne West laboratory, on a visit in the summer of 1993.

The outlook was strikingly similar to Rannoch Moor in the highlands of my native Scotland. Except that the tiny white specks dotted across the landscape were not isolated crofts, but more than 50 experimental nuclear reactors constructed, and in most cases discarded, during the Cold War.

The sheer scale of the Idaho National Engineering and Environmental Laboratory — a somewhat artificial entity dreamt up by Washington bureaucrats to manage and clean up these reactors — would overwhelm any visitor recently arrived from Europe, where nuclear technology has been developed in relatively cramped conditions.

The geographical, technical, human and financial resources available to the United States when it developed the atomic bomb, and still available to it today, struck home with me on that summer's day in Idaho.

And the sight of a vast convoy of cars and yellow buses crawling across the moor, returning thousands of laboratory workers to their Idaho City homes at four o'clock on a Friday afternoon, made a similarly lasting impression.

Every state in the union gets its share of the United States' technical and budgetary largesse, thanks chiefly to the composition of the US Senate, which has two powerful senators for each state, large or small (try that in

Europe, and see what it does for science in Ireland or Portugal). In the case of Idaho (population 1 million) half a billion dollars and ten thousand well-paid jobs, watching over part of America's nuclear legacy, will do nicely.

When Ted Steven (Republican, Alaska) took over from Mark Hatfield (Republican, Pennsylvania) — a strong ally of biomedical research — as chairman of the Senate Appropriations Committee, a concerned friend of the National Institutes of Health fretted that Alaska didn't have single medical school of note. "It will have soon!" retorted a more optimistic colleague.

For, although everyone is in favour of funding 'the best science', the fact of the matter is that the much-maligned pork-barrel politics is fundamental to the breadth of excellence and opportunity that characterizes the US research system.

Changing the guard

That system distributes and nurtures excellence through a peculiar mixture of honest peer-review and crude political power-broking, administered through a fairly chaotic network of rival agencies.

Unlike the more bureaucratic systems of Europe and Japan, the system allows for projects with which the Congress loses patience to be speedily abandoned — including umpteen of the Idaho research reactors and, more famously, the Superconducting Super Collider, whose death in 1993 remains the defining moment in US research policy over the past decade.

The significance of that decision lay chiefly in its psychological impact on the nation's physicists, whose position as the élite cadre of researchers had already been undermined by the passing of the Cold War. The 1990s has witnessed a changing of the guard, and the emergence of a new élite of life scientists, whose probing of the nature of life



ARGONNE NATIONAL LABORATORY

Roll out the pork barrel: federal largesse gave Idaho the Argonne National Laboratory-West.

captivates a public now less than entranced by further exploration the atom.

Their new pre-eminence will be confirmed when a future president of the United States appoints a biologist as his or her science advisor, a position that has always been held by a physicist. The job's importance seems to have diminished in recent years — which is odd, given that US politicians and their advisers have newly fallen in love with basic scientific research.

Bill Clinton came into office in 1993 insisting that technology — as opposed to science — could revive America. The aura of Japan's Ministry of International Trade and Industry was still intact at the time, and he proposed pouring billions of dollars into technology programmes intended to revitalize American industry. Eight years later, US industry has revived itself, and conventional wisdom holds that the United States' lead in basic science has contributed greatly to the current economic boom.

Crawling towards 2000

But the new conventional wisdom may have limited currency. Many of the most successful 'high-technology' businesses, ranging from America Online to the strange collection of wannabe Internet retailers whose television advertising has dominated the airwaves in the run-up to the holidays, have little to do with either science or technology.

These peculiar television ads, alternating with equally desperate news bulletins warning the American population of hypothetical millennial terrorist threats, provide an unap-

pealing background noise for the greatest power the world has ever seen, as it crawls furtively into the year 2000.

It was ever thus: as Harold Evans, the English journalist, points out in a recent edition of *US News and World Report*, similar apprehensions were widely expressed at the dawn of what turned out to be the American Century. For that we should perhaps be grateful. Despite the opposite impression prevalent overseas, a general absence of hubris is one of America's abiding strengths.

To give one tiny example, Americans — from talkshow host Jay Leno down — believe that their schools don't work, partly on account of the widely publicized Third International Mathematics and Science Study, published in 1998. Yet Jon Miller of the

Chicago Academy of Sciences has shown that US adult understanding of science is among the very highest in the world, owing to the enormous number of Americans who learn more science at college.

Of course, this doesn't mean that sound science steers US public policy. Take global warming: seven years of data, not to mention generally mild weather, have made no impression whatsoever on Congress's obstinate refusal to face the problem.

The United States, in that time, has come a long way towards recognizing the interplay between science and the economy. It is accepted that science can help to make the country both rich and healthy. Whether it can also make it more enlightened is an open question.

Colin Macilwain

'The role of science is to illuminate political choices, not enforce them'



Declan Butler is European correspondent for *Nature*, and has been based in Paris since 1993.

Paris

"I must reveal that in an earlier incarnation many years ago, I earned my living by writing for *Nature*, the world's most venerable scientific journal", wrote the anonymous columnist 'Bagehot' in May. He was delivering a typically provocative and pertinent analysis in *The Economist* on the British government's regulation of genetically modified organisms (GMOs).

The experience, he continued, "gave me a damaging regard for scientists", and he concluded that "simply quoting scientific authority is no answer to the conundrum of public trust". Rather, this could only be achieved by transparent, impartial decision making based on wide consultation.

After eight years of writing for this venerable journal, I do not have a damaging regard for scientists, far from it. But a recurring theme of my experience of reporting on issues such as France's blood scandals, human cloning, bovine spongiform encephalopathy (BSE), xenotransplantation and GMOs, is that the role of science — with all its attendant uncertainties — must be to illuminate political choices, not enforce them.

The risk of over-dependence on experts has been illustrated ad nauseam in our news pages. An epidemic of Creutzfeldt-Jakob disease of unknown proportions is hanging

over the United Kingdom and the many other countries whose citizens ate British beef contaminated with the agent that causes BSE.

People throughout the world have been infected with HIV by transfusions of contaminated blood in the mid-1980s. In both cases, experts were not only fallible, but were all too often swayed by economic and political considerations.

Backlash

The backlash against experts has been most dramatic in France, where several individuals have been sent to prison in

connection with contaminated blood supplies. Britain has taken a more dispassionate stance on its BSE crisis, convening a long and thorough £16 million (US\$26 million) inquiry to determine eventual responsibility. The inquiry's conclusions are likely to reinforce Winston Churchill's maxim that "scientists should be on tap, not on top".

But these backlashes are part of a wider change. We are in the middle of a profound and irreversible restructuring in the contract between scientists and society that has been in place since the Second World War (see *Nature* 402 supp, C81; 1999). That is why it is telling that even *The Economist*, the most influential mouthpiece for free trade and technological progress, acknowledges the need for a redefinition of the limits of scientific authority.

For one thing, under international trade law, any country refusing to import a product on safety grounds must justify its action. But a compromise now seems inevitable between free trade and demands to refuse imports on the basis of scientific uncertainty — such as the European Union's rejection of US hormone-treated beef, or France's refusal to respect the lifting of the European embargo on British beef.

Yet the phenomenon of 'globalization' has been ignored too long in such debates. Classical risk-assessment procedures, for example, largely fail to take into account the speed and scale with which new technologies are introduced worldwide — the relatively new risk is that if something goes wrong, it will go wrong in a big way.

Debates over GMOs and xenotransplantation show that, ultimately, the public's acceptance of new technology has less to do with science than with an



Up in flames: the BSE crisis illustrated the risk of over-dependence on experts. The public's trust in science and scientists, in Britain and elsewhere, has suffered as a consequence.

GREG WILLIAMS/REX FEATURES

assessment of the global balance of power among the major players in the debate, and its implications for the legitimacy of the scientific arguments being put forward.

The estimated \$6 billion market for xenotransplants, for example, and the huge public demand for organs, have been intelligently interpreted by many as considerations that could influence assessment of the real risk that the technology could create pandemics (see *Nature* 391, 320–325; 1998).

It has been of little help that the most outspoken scientists on the matter tend to be those with interests in seeing the technology progress. The delays to introducing the technology may not have pleased private investors. But they have forced a broader discussion of the issues.

As for GMOs, a massive campaign by Monsanto to educate the public only reinforced perceptions that the company was unhealthily powerful and keen to flatten debate. The public needs to be reassured

that the balance of power on such issues — long dominated by powerful multinationals and the scientific community — has become more even and trustworthy.

A new contract is evolving, in which scientists, like other 'experts', will be forced to redefine their roles in shaping public policy. The violent backlash against experts, and current resistance to GMOs, can be seen as an inevitable part of a transitional period from the old order — where scientific authority prevailed — to a more sophisticated approach to managing scientific and technological change.

Some common threads are already emerging, such as the need to broaden the expertise of advisory committees to include consumer and other organizations, complete transparency in decision-making procedures, and not allowing wolves (such as agricultural ministries) to guard sheep (such as public health). In this new order, scientists and their professional societies will need to be more active in carving out a

role as honest brokers who can help clarify the issues and ensure impartial information.

But this rethink could not have come at a better time. Over the next few decades, the human and environmental consequences of scientific research will become much broader, and the lessons of the last decade will need to be taken on board if the full promise of biology is to be realized.

This may require scientists taking more of a back seat. Human genetics is leading to a worrying rise in many scientific quarters of insidious and naive genetic determinism, which, if allowed to dominate public policy, could ultimately discriminate, ostracize or even eliminate those not meeting some genetic norm.

Thomas Jefferson did not need genetics to know that humans are not born equal, and that constitutions should compensate for these inequalities. Many of the ethical issues raised by genetics are human rights issues. Scientists are not experts in this area — and thus have little to say. **Declan Butler**

'Japan's scientists must learn to take action'



David Swinbanks is the Asia-Pacific editor of *Nature*, and has been on its Tokyo staff since 1985.

Tokyo

At the start of the new millennium, can scientists in Japan look forward to a bright future? On the surface, things seem to be improving. Following the new basic law for science and technology and the associated five-year plan, launched in 1995, large amounts of extra public funds have been pumped into science.

Japanese scientists are publishing increasing numbers of papers in international journals, and Japan stands second only to the United States in terms of the output of papers in the Science Citation Index of the US Institute of Scientific Information. But underlying these superficial improvements, deep-rooted problems remain.

Japanese scientists made a significant contribution to the recent milestone paper on the sequencing of human chromosome 22 (*Nature* 402, 489–495; 1999). But while their counterparts in the West worked in comfortable, modern facilities armed with bankloads of DNA sequencers, the scientists at Keio University, who made the Japanese contribution, had to make do with a handful of sequencers in a dingy, cramped laboratory in the basement of one of the university's oldest buildings.

The appalling laboratory conditions in

Japan's leading universities were highlighted in the early 1990s by Akito Arima, then president of the University of Tokyo. In response, the government pumped money into new university buildings. But most university researchers must still make do with buildings and laboratory conditions more akin to those in the Third World than in one of the world's leading economies.

To some extent, conditions have actually worsened. In a government push to strengthen graduate schools and university research, the number of graduate students in national universities has doubled in the past decade to 180,000. But the laboratory floor space available has increased by only 10 per cent, according to a recent report by the Japan Science Council, a body of academics that advises the government. Thus, researchers are crammed into even less space.

Starved of funds

The five-year plan to increase public spending on science by 50 per cent between 1996 and 2001 has triggered a plethora of large research grants from the various science-related ministries. But most of this extra money has gone into buying new equipment rather than the fundamental renovation of infrastructure that is required.

Furthermore, the money is dispersed in an uncoordinated and wasteful fashion by the several ministries, with some researchers receiving far more than they need while others who may be equally deserving are starved of necessary funds.

This is the case with the researchers at

Keio University who, despite making world headlines, are not benefiting from the bonanza in funding for genomics-related research that the government plans to provide at the start of the new millennium (see *Nature* 402, 569; 1999).

Who is to blame? Much of the fault clearly lies with the fact that the ministries and agencies responsible for science are more concerned with preserving and expanding their own budgets than with focusing on the overall needs of the country.

The merger of the Science and Technology Agency with the Ministry of Education, Science, Sports and Culture (Monbusho), due to be complete by next year, is an attempt to restructure and improve efficiency in the government's management of science. But the officials at the education ministry are experts at creating inefficient and ineffectual bureaucracies, and are hardly in a position to lead restructuring forward.

But Japan's scientists are also to blame for their passive acceptance of the poor conditions under which they have to work. Reformers such as Arima who lobby for change are the exception, and, rather than receiving support from their colleagues, they are often criticized for being too outspoken.

The university system is riddled with unproductive faculty members content to work out their lifetime employment with the minimum of effort. As a result of the Asian recession, Japanese industry is in the throes of major restructuring.

A similar shake-up of the university system is required. The government's moves to

convert universities into semi-autonomous 'agencies' is being presented as an attempt to introduce more efficiency, accountability and cost-effectiveness. But the real drive behind this proposal is merely a shallow attempt by politicians to reduce, on paper, the number of government employees by removing university staff from the official government payroll, while taxpayers will continue to foot the bills.

Japan's best scientists will no doubt soldier on, making significant contributions to science in the face of adversity. Their situation, however, will only improve if scientists themselves join together to form more effective bodies to lobby the government and to build greater public awareness and under-

standing of science. They also need to make far greater efforts to bring about much-needed reform within their own institutions.

Some small steps in this direction have been taken. Early last year, for example, Earth, marine and atmospheric scientists from various departments at the University of Tokyo linked up to create a new interdisciplinary regrouping of departments. The new arrangements were then opened to external comment and review by a team of foreign and Japanese scientists.

Only through such grassroots activity by scientists themselves can there be hope for significant improvement in the working environment of Japanese scientists in the new millennium.

David Swinbanks

'I have reported on astonishing achievements in times of hardship'



Alison Abbott is senior European correspondent for *Nature*, and has been based in Munich since 1992.

Munich

I seem to have spent much of my seven years as *Nature*'s Munich correspondent reporting depressing news of recession, budget cuts, broken political promises and disappointed scientific hopes.

Certainly all the European science organizations, including the European Southern Observatory (ESO), the European Laboratory for Particle Physics (CERN), the European Molecular Biology Laboratory and

the European Space Agency (ESA) have seen their budgets fall in real terms during these years. And the erosion of research budgets in Germany, Italy and neighbouring countries have confirmed the fickleness of political promises.

But I have also reported on astonishing achievements in times of hardship. For example, CERN is building the Large Hadron Collider, which will be the world's most powerful particle accelerator when it begins the search for the Higgs boson in 2005.

ESO's Very Large Telescope, now being built atop a remote mountain in northern Chile, will be the world's largest and most advanced optical telescope when it becomes fully operational in 2003. ESA, despite its clipped wings, has launched several important scientific missions, including the X-ray satellite XMM launched just before the end of the millennium. Unfortunately, no such grand visions are pushing the frontiers

in the life sciences, where investment is sadly lacking both in Germany and Italy.

But the greatest achievement that I have witnessed has been the rebuilding of research in east Germany after reunification (see *Nature* 401, 635; 1999). This has been more than just watching buildings sprout up among the *Plattenbau* (concrete apartment and office blocks). It has also been about watching the complex social readjustments within the east German scientific community, which had lost so much during the last ideologically and financially bankrupt years of communism.

Many scientists lost even more following reunification. More than half lost their jobs,

and most of those who did not found themselves working for west German research directors. Few could seek refuge in a sense of injustice, as personal restrictions and material shortages were immediately replaced by freedom and plenty, courtesy of the west. Even those most deeply humiliated by the west German take-over admit to its historical inevitability, even necessity.

Interviewing in east Germany has always been a poignant experience, leaving feelings of both optimism and sadness, and a vague guilt for my luck at having been born in peacetime and on the right side of the Iron Curtain. The saddest story I covered concerned the suicide of an east German professor who was cleared of abusing communist party connections to the detriment of his colleagues — but only after his job had been given to someone else.

The most frustrating stories to unravel have been those on the restructuring of a reunited Berlin, for which the western side of the city paid a high price. It was hard to feel too sorry for west Berlin scientists, who had been kept in relative luxury — sometimes indolence — by subsidies that vanished when the city lost its island status, and who were forced to share their bounty with their new colleagues in the east.

Interviewing in Berlin was exhausting, as both sides of the divide expressed their deep frustration, or even anger, with gusto, wearing their mutual antipathy and lack of understanding with pride.

In the relative calm that now prevails, it is clear that east Germany has gained much more than it has lost from post-communist restructuring. So has west Germany, in that cash shortages have forced it to become more competitive.

Further east, the balance is less clear-cut.



Walls came tumbling down: the reunification of Germany led to a short, sharp shock for east German scientists, but the whole country's science has emerged the better for it.

Central and eastern European countries, waiting hopefully in line for membership of the European Union, have been through much the same as east Germany. The structures of the old academies have been replaced, for example, and new university laws have been approved.

But without the cash injection that east Germany enjoyed, progress has been slow. Moreover, the passive mentality bequeathed by decades of communism seems to be harder to shift without the short, sharp shock experienced in east Germany.

Most depressing has been the failure of Italy, with its long and admirable scientific tradition, to reform its scientific structures to curb the power of the *baroni*, as powerful professors are known, and let a competitive and meritocratic system flourish.

Opportunities that opened up with the demise of the corrupt Christian Democrat governments in the mid-1990s were grabbed only half-heartedly, and the mystery of Italy's introspective stagnation continues, along with the mystery of its occasional successes.

Alison Abbott

'If knowledge is king, we may need a republican revolution'



David Dickson has been news editor of *Nature* since 1993, and was its Washington correspondent in 1978–82.

London

Last month, more than 300 years after the philosopher Francis Bacon coined the phrase 'knowledge is power', a British cabinet minister came up with an even stronger aphorism. In the years ahead, he said, "knowledge will be king".

The steady growth in the importance of science over the past decade — and the unprecedented growth in science budgets over the past 50 years — confirms this statement, and the way in which it sets a theme for the next millennium.

Some scientists will, no doubt, continue to claim well into the next century that they are failing to receive the recognition they deserve. But, with politicians and economists committed to the expansion of a global knowledge economy, there is ample evidence that any remaining financial restrictions are more the result of economic constraints than a lack of political (or popular) will.

Yet the 'regal' power bestowed on science has its dangers. Some, as Declan Butler points out above (see page 6), feel that the authority it appears to provide — particularly to those grasping for platforms of apparent certainty in an uncertain world — is misguided at best and tragic at worst.

One constant theme of the news pages of *Nature* over the past decade has been the unedifying sight of politicians trying to wriggle off the hook on which they have impaled themselves by a commitment to the 'guaranteed' safety of processes and products. Examples range from the storage of nuclear waste to beef contaminated with bovine spongiform encephalopathy.

Scientists often distance themselves from such guarantees; they are the first to recognize that scientific 'truth' is approximate and transitory. The problem has come not from the respect that scientific knowledge deserves, but from the 'absolute' authority it can be given in a political context.

Or take the issue of knowledge as intellectual property. Before coming to *Nature*, I had little idea of what a patent was, and even less of its implications. Since then, I have covered stories ranging from the Supreme Court's key decision in the early 1980s that life can be patented, to recent disputes over the rights to the enzyme *Taq* polymerase.

A glance at this week's news pages, with separate stories about patents on high-temperature superconductors, on plants and on genes, will confirm how far science's evolution from public knowledge into an essentially private commodity has penetrated to the core of the scientific enterprise.

The dominant theme of this coverage is not the legitimacy of the patent system, acknowledged in the US constitution as an appropriate way of rewarding inventors,

including scientists. Rather it is the way in which the power given to an inventor is used, and the danger of it being used in an absolutist fashion.

The rights to life

In principle, patents were meant to ensure that individuals shared their inventions for an appropriate reward, and in most cases this is what happens. But much of our news coverage has, by its nature, highlighted allegations of this reward being wrongly claimed (as in the recent court disputes over *Taq* or human growth hormone), too greedily pursued or too aggressively exercised.

Directly related to this is the increasing reluctance of scientists to share data on the grounds that it is potentially profitable — but only if patented. A footnote to a recent press release from a leading US genome-sequencing company puts it eloquently when it warns that one risk to its business prospects is the "adverse effect of public disclosure of genomic sequence data".

There is, of course, no room for populism in science — the spectre of creationism shows this. Nor can science be democratic in the political sense. The quality of a scientific idea is not measured by the number of votes it is able to gather, even in the scientific community; it can only be judged by a rigorous peer-review process with its own tested rules of procedure.

But the authority that this process gives to science can be used or abused. Used responsibly, it offers the prospect of a healthier, better-fed and more prosperous world — hopefully the prospect opening before us. Used irresponsibly, it can increase the control of the powerful over the powerless, and widen the gap between haves and the have-nots.

If science has, indeed, become 'king', it may be time for a truly republican revolution. And perhaps the opening of a new millennium is an appropriate setting for such an event.

David Dickson



Private property? Protesting against the patenting of cloned animals.

NIH guidelines urge scientists to share research tools

Washington The US National Institutes of Health (NIH) has published guidelines that actively encourage researchers to share NIH-funded research tools that they create. But it has rejected suggestions that it should be able to lay down the terms under which public- and private-sector researchers offer access to such research tools.

The new guidelines, which were published two days before Christmas, had been requested by Harold Varmus, the outgoing NIH director. A draft version was published for comment earlier this year (see *Nature* **399**, 291; 1999).

One of the aims had been to limit the extent to which the inventors of basic research tools claim 'reach-through' rights. The guidelines say that imposing this condition of using an NIH-funded research tool is "inconsistent" with the principle of open access on which they are based. But they stop short of forbidding reach-through entirely, saying that NIH grant recipients will have "considerable discretion" to interpret the guidelines.

The principles on which the guidelines are based are directly applicable only to

recipients of NIH funding. But the NIH hopes that other not-for-profit and for-profit organizations will adopt similar policies and refrain from seeking unreasonable restrictions or conditions when sharing materials. The guidelines are available at http://www.nih.gov/od/ott/RTguide_final.htm

Portugal bans planting of genetically modified maize

Barcelona Portugal last week banned the planting of genetically modified (GM) maize (corn) marketed by Compa and Elgina, the local affiliates of Novartis and Monsanto, because of "concern about potential environmental risks of large-scale planting".

Last February, the Ministry of Agriculture authorized the production of GM maize following the completion of field tests. There are currently 1,300 hectares of GM maize in Portugal, representing 0.5 per cent of the country's total crop.

Japan restricts blood from visitors to Britain

Tokyo Japan's health ministry is to restrict the donation of blood from people who have spent time in Britain, in an attempt

to minimize the theoretical risk of introducing new-variant Creutzfeldt-Jakob disease (vCJD).

According to an announcement by the ministry on 25 December, Japan Red Cross, which is responsible for collecting blood in Japan, will no longer accept blood from people who lived in the United Kingdom for more than 6 months between 1980 and 1996. The ministry says it is keen not to put off potential blood donors or cause concern among those in this category, but feels the measure is needed to eliminate potential risk factors until more research is carried out on vCJD.

Europe in front on list of number of papers published

Paris Europe published more scientific papers than the United States in 1997 for the first time since the Second World War, according to a report produced by the Paris-based Observatoire des Sciences et des Techniques. The study says that the European Union published 33.5 per cent of the total, compared with 32.6 per cent from the United States.

The report, *Indicators 2000*, says that the shift is partly due to an increase in scientific publications from Austria, France, Ireland, Italy, Spain and the Scandinavian countries. Germany's contribution stayed

level at 6.6 per cent of the total between 1982 and 1997, while that of the United Kingdom fell from 9 to 8.4 per cent.

Motorola reveals plans for centre near MIT

Boston The communications company Motorola has announced the creation of a \$5 million Consumer Design Center near the Massachusetts Institute of Technology (MIT) campus in Cambridge, Massachusetts. Scheduled to open in mid-February, the centre is intended to serve as a “catalyst for the development of innovative technology solutions that will change people’s lives”.

Motorola hopes to develop products in a number of areas, including wearable, voice-controlled computers, wireless Internet communications devices, and ‘Dick Tracy’-style wristwatch phones. The company has worked with the MIT Media Laboratory since 1994, sponsoring a fellowship programme that fosters collaboration between the company and researchers in the laboratory.

Japanese panel to discuss uses of human DNA

Tokyo Japan’s Science and Technology Council, its highest-ranking science

policy-making body, has created an ad hoc committee to discuss the ethical and social implications of research using human DNA samples. The committee will meet later this month to begin discussions on issues such as protecting personal genetic data and commercializing genetic research. It plans to compile a draft report by the end of March.

The Ministry of Health and Welfare set up a working group last year to create ethical guidelines for the management and use of human genetic information (see *Nature* 402, 8; 1999), but this concerns only medical institutions overseen by the ministry. The council now hopes to create a policy outline that can be applied on a national level.

US eases sanctions on Indian laboratories

New Delhi The United States has agreed to lift the sanctions on 51 Indian laboratories and industrial centres that were imposed after the nuclear tests conducted by India in May 1998. The move will enable scientists in these laboratories to obtain US goods and technology, whose shortage had slowed down research (see *Nature* 397, 460; 1999). The sanctions had also effectively barred them from visiting US laboratories.

The decision, announced on 17 December, follows a congressional directive to reduce the sanctions list of 204 Indian institutions to “those entities most directly involved in proliferation activities of concern”. The main beneficiaries will be basic research institutions under the Department of Atomic Energy, such as the Tata Institute of Fundamental Research (TIFR) in Mumbai and the Institute of Physics in Bhubaneswar.

US spending on R&D rises by 71 per cent

Washington US businesses have boosted their investment in research and development (R&D) from \$97.1 billion in 1994 to an estimated \$166 billion in 1999 — an increase of 71 per cent — according to a report from the Industrial Research Institute of Washington.

Most of the increase comes in the life sciences and information sciences, including telecommunications. John Jankowski, a financial analyst at the National Science Foundation, says that the estimate is “right in line” with the foundation’s own figures. The US agency will publish actual numbers of private R&D investments from 1998 later this month.

Do we need a huge new centre to annotate the human genome?

Sir — Now that a human chromosome has been sequenced¹, and a rough draft of the complete genome promised for the spring², one obvious requirement is to annotate the genome; that is, to precisely locate all the genes and assign them to their protein products. There are sure to be dozens of applications for “human genome annotation centres”, all emphasizing the strategic nature of this endeavour and the huge expected benefits for national biomedical industries. How can we avoid a planetary-scale duplication of effort caused by the creation of multiple human genome databases, all using different protocols, software and format?

Paradoxically, the high number of candidate centres arises because we do not yet have the bioinformatic capacity to annotate 3 gigabases of human genome sequence accurately. Although the best gene-finding programs can detect protein-coding exons with 80% accuracy³, performances on identifying 5' and 3' untranslated regions, core promoters or other regulatory sequences are close to zero. Moreover, alternative transcript forms (probably involving more than 30% of genes) are not addressed by any program, and non-protein-coding genes (for example *XIST*, *IPW* or *NTT*) are totally transparent to non-similarity-based detection programs³.

A recent complementary-DNA-based survey of *Caenorhabditis elegans* (nematode worm) annotations reported by Jean Thierry-Mieg⁴ shows that only 50% of proteins are correctly predicted, the rest exhibiting various degrees of annotation errors (for example, incorrect, added or missed exons; fused genes; or being entirely wrong). The *C. elegans* genome is more compact, much smaller and much less repetitive than the human. An impressive gathering of leading computer biologists has therefore recommended the sequencing of more expressed sequence tags (ESTs) and cDNAs as the best way to annotate the human genome⁴.

In the absence of accurate automated methods, the level we can hope to achieve will rely on a fairly trivial protocol, mostly based on similarity searches, which could be set up by any competent bioinformatics laboratory within weeks. The cost of the required computing power is well under \$300,000, if a PC cluster solution is retained. For instance, a cluster of 125 Pentium 3 (see “gigablast” at <http://igs-server.cnrs-mrs.fr>) would take one day to compare all human EST contigs to the whole human genome sequence, and the rest of the week to compare it with all known proteins.

However, the simplicity of protocols like this one is also a recipe for disaster. At each step, a variety of programs, arbitrary parameters and thresholds can be chosen. Multiple, independent genomic annotation efforts would thus result in highly redundant human genome annotation databases.

Going further than the above automated annotation stages will require a lot more money and a bigger organization. For example, detailed visual inspection of each gene will be needed to reduce the 50% error rate significantly. I estimate the workload at one gene each day per annotator: 300 person-years in total. Starting from a unique, single-pass, automatically annotated human genome sequence, such work could be split among 30 international groups of five people each. This would complete the work in two years.

What about the “strategic value” of an annotated human genome sequence? Does it justify launching multiple efforts?

I think not, for two main reasons. First, most pharmaceutical companies seriously involved in genomics have already identified many targets using cDNA and EST data from private sources. It is unlikely that many new genes of economic value will be revealed by the systematic annotation of the human genome (especially if public EST and cDNA data are central in the protocol).

Second, increasing our knowledge of human genomic sequence at a large scale is sure to generate significant improvements in the next generation of gene-finding software. It would then be ridiculous to have embarked on a detailed annotation effort using obsolete methods.

Jean-Michel Claverie

Structural & Genetic Information Laboratory,
CNRS-UMR 1889, Marseille, France

1. Dunham, J. et al. *Nature* **402**, 489–495 (1999).
2. Wadman, M. *Nature* **398**, 177 (1999).
3. Claverie, J.-M. *Hum. Mol. Genet.* **8**, 1821–1832 (1999).
4. Cold Spring Harbor First Workshop on Computational Biology: Bridging the Gap between Sequence and Function (7–9 September 1999).

How nature itself uses genetic modification

Sir — Mae Wan Ho¹ states that genetic engineering is fundamentally different from conventional plant breeding or wide crosses to produce novel crops; thus she concludes that genetically modified (GM) plants need unique tests on their safety.

Any new crop should have its composition of major and minor constituents investigated to establish substantial equivalence, and its novel trait independently tested, as has been the case for the present range of GM plants. However, Ho states that special tests are required because “genetic engineering enables exotic genes... [to be]

combined in novel constructs, often with viral promoters to make genes overexpress continuously. The constructs are inserted into genomes by transformation techniques that cannot control where the genes go, resulting in a range of unpredictable positional effects and rearrangements”².

But this differs little, if at all, from Ho's description of normal genome behaviour in her book², which states: “Genome organisation is infinitely variable [and contains] transposons that can excise and reinsert elements in different locations in the genome” and “up to 20% of some genomes may contain reverse transcripts. These processes destabilise genes and genomes, move genes around, mutate, rearrange, recombine, replicate sequences...”. What could be more exotic than a mutant protein?

Some estimates suggest that the genome of some cereals may contain up to 50% retrotransposons; transposons contain end regions that are hot spots for recombination using transposase. The plant genome contains very large numbers of strong promoters that direct expression as strongly as any viral promoter such as the cauliflower mosaic virus 35S promoter. These promoters are used experimentally; many of them are constitutively and continuously expressed because they control the expression of housekeeping genes. We know the consequences of events via random GM insertion from libraries constructed from T-DNA insertion. However, Ho and other critics of GM technology neglect the extent to which selection is made among many GM transformants, as is the case with conventional plant breeding among siblings. Lethal insertions are self-selecting; potentially innocuous insertions are detected by substantial equivalence.

It is important to recognize that all the food we eat has been (and is) continuously genetically engineered by natural phenomena in ways that do not differ in any fundamental way from the current GM technology. Natural genetic modification of wheat³ and rice, for example, enabled the breeding of dwarf crops, used to feed many millions in the ‘Green Revolution’.

To criticize experimental GM technology while accepting the benefits of natural GM implicit in ‘the fluid genome’² hints at a not uncommon attitude that sees synthetic pesticides as morally reprehensible but natural pesticides as good. Is this really any different from notions of Original Sin?

Anthony Trewavas

Institute of Cell and Molecular Biology,
University of Edinburgh, Edinburgh EH9 3JH, UK

Christopher Leaver

Department of Plant Sciences, University of Oxford,
South Parks Road, Oxford OX1 3RB, UK

1. Ho, M.-W. *Nature* **402**, 575 (1999).
2. Ho, M.-W. *Genetic Engineering: Dream or Nightmare?* (Gateway, Bath, 1998).
3. Peng, J. et al. *Nature* **400**, 256–261 (1999).

Millennial highlights

... from Gerbert d'Aurillac to Watson and Crick.

J. L. Heilbron and W. F. Bynum

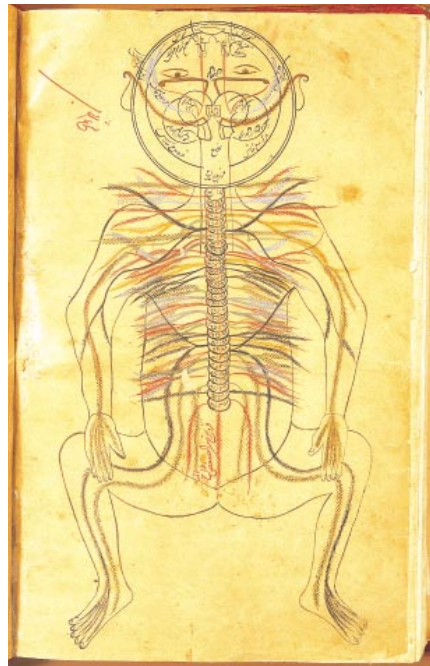
That grand insight of the twentieth century, that science is a social process, has penetrated even into the black box where we fashion our masterpieces. It has inspired us to replace our usual celebration of great and near-great individuals with snapshots of the state of science and learning around century and half-century dates, beginning in the year 1000. To give an impression of accelerating progress, we take our first snapshots at quarter-millennial intervals (1000, 1250, 1500), slow down to the standard unit of a century (1600, 1700, 1800), and end with clichés every 50 years (1850, 1900, 1950).

In order to impose some discipline on our fancy, we draw our examples from a period of seven years on either side of half-century dates, and 14 years on each side of century dates. The facts have forced us to allow a margin of 28 years around 1250, which is our only intermediate point in the demi-millennium 1000–1500. The significance of the number seven and its multiples in anniversary computations has not been recognized previously. To indicate that our century dates have brackets, we introduce the convention that '1X00' signifies the century date 1X00 together with its appropriate margin of seven; for example, '1500' means 1500 ± 14 .

For each '1X00' we call attention to one or more synthetic works or compendia created during the interval. It is noteworthy that Copernicus's first draft of his world system (1514), but not his final account of 1543, and nothing by Descartes, are eligible. As for Galileo, our method singles out his *Sidereus Nuncius* (1610) but does not catch his condemned *Dialogue on the Two Chief World Systems*. We are able to highlight Darwin's work on barnacles but not *On the Origin of Species*. Whether these omissions indicate a fault in our method or the need to revise the Canon of Science is under investigation.

The first millennium (1000 – 14)

Around 997, Gerbert d'Aurillac, Bishop of Rheims, soon to be Pope Sylvester II, wrote to a former student about the then challenging question of the area A of an equilateral triangle of side a . The manuals of the *agrimensores* (surveyors) gave $A = a(a + 1)/2$. Gerbert observed that the formula should be $3a^2/7$ or $13a^2/30$; an improvement not worth mentioning were it not that Gerbert and his school represented the highest level of mathematics in the Latin West at the turn of the first millennium. By then, Arab geometers



Setting the pace: it took Western scholars 400 years to catch up with Avicenna's *Canon*.

and astronomers had long since mastered and gone beyond the technical treatises of the Greeks, and had incorporated this knowledge in beautiful brass astrolabes with which they could solve difficult problems of spherical trigonometry. An exemplary cultivator of their mathematical way was Ibn al-Haytham, or Alhazen (965–c. 1040), author of, among much else, a book on optics that became the standard guide to the theory of light and vision in the mediaeval West.

Around '1000' the West was just beginning to make fruitful contact with Muslim centres of learning. Gerbert himself studied in Spain, where he perhaps became acquainted with Arabic science. The treatise *De Astrolabia* is sometimes attributed to him, and also an astrolabe, which has been preserved. In any case, his reputation as a mathematician and position in the church indicates the keen interest in geometry and astronomy among learned Europeans around the year 1000, and the distance to be travelled before the experts in the West could match the savants of the East.

The precocious philosopher and physician Ibn Sina (Avicenna), reputed to have begun the practice of medicine at the age of 16, in 996, may be taken as the emblem of Muslim learning. His bibliography numbers 270 items, including the five-volume, encyclopaedic *Canon of Medicine*, which consolidated the combined medical knowledge of

the Greeks, Romans and Arabs. It took physicians in the West over four centuries to digest its million words, and some medical schools in the Middle East are still chewing them over.

Our choices of compendia for '1000' are Avicenna's *Canon* and the *Optics* of Alhazen.

Aristotle's ups and downs (1250 – 28)

By the middle of the thirteenth century the natural philosophy of Aristotle and the great Greek mathematical books — Euclid's *Elements*, Ptolemy's *Almagest*, some of Archimedes and Apollonius — were available in Latin translation in the major universities of the West. Epitomes and compilations, lectures and disputations brought ancient science to within the grasp of students in the new arts faculties. Some regent masters wrote extensive commentaries on Aristotle's *libri naturales* and a few went beyond what they received. Thus the Dominican Saints Albert the Great and Thomas Aquinas accommodated Aristotle and Christianity from the 1240s to the 1270s; the Pole Witelaw and the Franciscan Roger Bacon improved on Alhazen with the help of experiments (c. 1270); and several clerics proposed calendar reform along the lines adopted much later by Pope Gregory XIII.

Aristotle's nature books, banned at Paris early in the thirteenth century, had become the basis of science teaching there by the mid-century. The ban was lifted in 1234, after provincial universities had attracted students away from the capital by allowing the reading of the Master of Those Who Knew. Unfortunately, the Master had not had the happiness to know the truths of revelation, which some of his best-founded principles violated. For example, Aristotle proved almost conclusively that nature does not allow a vacuum; and some of his latter-day disciples in Paris transmitted this proof without hinting that God Almighty could make a void if He wanted to. In 1277 the Bishop of Paris, Etienne Tempier, in his capacity as chancellor of the university, condemned many theses from the physics of Aristotle, some taught by Aquinas and Bacon, as subversive to faith or contrary to scripture. No doubt the bishop was right, as a matter of physics; which shows that the exercise of external authority against the freedom of the academy, though regarded, no doubt correctly, as anathema, can spot and perhaps curtail unhealthy fads among intellectuals.

From among the many *summae* composed during the interval, we pick Saint

Albert's *De Vegetabilibus et Plantis*, which contains the first European description of spinach, and that excellent brief guide to the advanced natural philosophy of the period, Bishop Tempier's list of condemnable propositions.

New worlds on Earth (1500 – 14)

"It was wonderful that Columbus discovered America. It would have been still more wonderful if he had missed it." Thus one of America's leading philosophers, Mark Twain, himself once a navigator on the mighty Mississippi, observed with a professional eye that when Columbus set his sails square to the trade winds, he would necessarily be blown into the Caribbean. Columbus touched down exactly when, but nowhere near where, he expected. He had estimated that his journey to "Cipangu", an island east of China extolled by Marco Polo, would take 30 days from the Canary Islands, an estimate grounded in gross and repeated misreading of the facts known to him and to the experts who advised Queen Isabella that the ocean was too big to cross.

In '1500' European adventurers rounded the Cape of Good Hope to gain India by sailing east (Vasco da Gama, 1488), discovered new worlds by trying to reach China by sailing west (Columbus, 1492–1504), explored the American coast from Brazil to New Foundland, established settlements in Mexico, and collected examples of plants, animals, minerals and diseases unknown in the Old World. Cartography, botany and zoology boomed. Descriptions of the New World spread far and wide with the help of the printing press; Montezuma's revenge spread just as efficiently with the help of the social life in seaports; and Europe began its unforgivable destruction of the native populations of the new worlds opening up before it.

Our publication of the demi-millennium is the 1507 map of the German cartographer Martin Waldseemüller, who first referred to the new continent as America.

New worlds in the heavens (1600–14)

Columbus died in 1506. Barely more than a century later, Galileo turned his telescope to the heavens. He found that Jupiter has moons, Venus has phases, the Sun has spots, and the Moon has mountains. Swarms of stars never before seen came into view. But whereas on the ground the Old World almost annihilated the New, in the sky the New World eradicated the Old; for, before the seventeenth century had expired, the ancient geocentric cosmos had been discarded in favour of a less flattering heliocentric picture that displaced humankind from the middle of everything to an undistinguished place on a minor planet on the periphery.

Several important Copernicans besides Galileo were active in '1600'. Johannes Kepler decreed his first two laws of planetary



The law-giver: Newton made physics inexorable and universal.

motion in his *Astronomia Nova* of 1609. René Descartes, whose analytical geometry and principles of philosophy would revolutionize mathematics and physics, graduated from his Jesuit college in 1612, sowed some wild oats in Paris, and, in 1614, sat down to work at mathematics. The doomed old way also prepared for the War of the Worlds: Cardinal (now Saint) Robert Bellarmine was perfecting the knowledge of astronomy and the skill at argument that climaxed in his admonition, as head of the Roman Inquisition, to Galileo not to teach Copernicanism. The future Pope Urban VIII, who would issue the fatal condemnation of Galileo in 1633, was developing his knowledge of the world, at least once in a friendly dispute with Galileo. The Jesuits, who may have played an important part in the condemnation of 1633, were busy confirming the observations that Galileo interpreted as proof of the Copernican world system.

And so we choose as the syntheses of the era Kepler's *Astronomia Nova* (1609), which provided new answers to old questions, and Galileo's *Sidereus Nuncius* (1610), which revealed things altogether new.

Law and order (1700 – 14)

The period running from the publication of Newton's *Principia* in 1687 to the first public lectures on and around his natural philosophy in London and Paris set many of the leit-motifs of Enlightenment science. The *Principia* provided a model for the exact description of nature, with the modern characteristic that few people understood it. By the time of its second edition in 1713, however, it was accepted as a *chef d'oeuvre* of mathematics, though not of physics, which at the time was still qualitative and, except for the new entertainment of experimental demonstrations,

largely bookish. It still included subject matter that now belongs to the biological and Earth sciences.

Physics began to lose both its bookishness and its life science under a new species of lecturer, who developed popular demonstrations of natural philosophy that privileged experimentation. Newton's second major work, *Opticks* (1704), offered much material for the lecturers and also, in the form of Queries appended to it, bold speculations about the nature of matter, light, electricity, and chemical phenomena for the armchair as well as the experimental philosopher.

Newton's work encouraged the idea that the physical world, including perhaps the human body and mind, operated more regularly and dependably than natural philosophers had thought likely. It was no longer a question of regularities that happened "always or for the most part" (as in Aristotle's physics) or that could be brought about in half a dozen different ways (as in Descartes'), but of universal laws, which, like gravity, act inexorably. Newton did not believe that the Solar System was altogether stable under universal gravitation, which, he reckoned, would gradually slow the planets down until they tumbled into the Sun; but that had the advantage of providing God with an opportunity for further employment, either topping up the slacking motions or deciding to allow the universal conflagration in which the world might end.

Newton's epigones John Freind and James Keill attempted to bring law and order to medical and psychological phenomena within a generation of the *Principia*'s publication. Organisms and minds have proved more refractory to Newtonian universalism.

Our syntheses of the interval, in first and second place, are Newton's *Principia* (1687, 1713) and *Opticks* (1704, 1706); a literary sweep that, together with his alchemy, makes him the *Harry Potter* of the history of science.

Imponderabilia (1800 – 14)

During the eighteenth century the increasing emphasis on experiment for exploration as well as demonstration in physical science created substantial subspecies of electricity, magnetism, heat and light. Toward the end of the century these fields were united methodologically by the scheme of imponderables, the Standard Model of the physical sciences around 1800. The scheme supposed the existence of weightless substances as the carriers of the forces supposed responsible for the various domains of phenomena. Electricity had two such substances, a positive fluid and a negative; magnetism two also, an austral and a boreal; and heat, light, and perhaps fire and phlogiston, one each.

In the more mathematical versions of the scheme, as developed by P. S. Laplace and his disciples, the imponderables acted on one another and on ponderable matter via forces

that decreased with some power of the distance. In other versions, the nature of the force or power carried by elements of the fluids was not specified quantitatively. The imponderables coupled easily to the new chemistry of Antoine Lavoisier via a heat able to combine with bodies, and to the new *biologie* of Jean-Baptiste Lamarck via the caloric, light and electricity that he speculated could fashion simple animals out of moist, gelatinous matter.

In '1800' the Standard Model of imponderables received strong confirmation from two sets of discoveries. One was the detection of radiant heat and the characterization of its relationship to light, particularly in the work of William Herschel and John Leslie. With this discovery the Standard Model possessed a set of analogies that ran from magnetism to electricity to heat to light, which suggested that the imponderables might be related substantively as well as methodologically.

The second confirmatory discovery resulted from the publication of Luigi Galvani's experiments on frog legs in 1790. Galvani invoked a new sort of electricity, an animal electricity, to explain his findings. He thereby brought into the Standard Model the fluids that many physiologists had imagined to fill the nerves and activate the muscles. In the subsequent dispute with Alessandro Volta, many phenomena concerning the "nerveo-electric fluid" and the electricity developed by the contact of dissimilar materials came to light. The discussion culminated in Volta's invention of the pile, or "artificial electric organ", the reference here being to the structure, on which Volta modelled the pile, by which electric fish accomplish their business.

With the continuous current produced by the pile, natural philosophers provoked chemical reactions, notably the separation of water into its constituents and the isolation by Humphry Davy of sodium and potassium from their salts. The evident connection of electricity with the forces of chemical affinity enlarged still further the putative domain of the Standard Model. Very influential new connections between the imponderables and ordinary matter occurred in John Dalton's ideas about chemical atoms enveloped in atmospheres of caloric and J. J. Berzelius's theory of the electrical bonding of chemical substances. The standard modellers of '1800' had every reason to be pleased with their construction — notwithstanding that soon enough it was shown to be altogether wrong.

Frenchmen, who pretended to decry the spirit of synthesis, are the authors of the compendia of '1800': Lavoisier's *Traité Élémentaire de Chimie* (1789), Laplace's *Système du Monde* (2 vols, 1796) and Lamarck's *Philosophie Zoologique* (2 vols, 1809).

Origins and prospects (1850 – 7)

By '1850', the sciences which the polymathic

Master of Trinity College, Cambridge, William Whewell, called palaeiological — dealing with past causes (and consequences) — had come of age. Mineralogy, geology and palaeontology had shown that the Earth's history was both vast and knowable. Even biology was becoming more historical. Whereas Lamarck, Erasmus Darwin and other late-eighteenth-century savants had failed to convince many of their fellow scientists (a word Whewell coined in 1833) that present-day animal species might be descended from older, extinct forms, historical thinking was now respectable in science. It fell to that most respectable of bourgeois, Charles Darwin, to provide the synthesis that still has currency. Darwin inherited a *Weltanschauung* in which the idea of progress was constitutive. He was nervous about its implicit teleology and tried unsuccessfully to expunge his writings of loaded words like 'higher' and 'lower'. As if to demonstrate that even despised creatures were worthy of loving scientific care, he devoted almost a decade of his life to studying extinct and extant barnacles.

While Darwin fretted about progress, two of his contemporaries uncovered bleakness. William Thomson and Rudolf Clausius were energetically looking at the irreversibility of heat transfer in that Victorian mainstay, the steam engine. There was, they calculated, always a price to pay, or as Clausius put it in Darwin's *annus mirabilis*, 1859, entropy inevitably increases. What this meant was

that, no matter how well Darwin explained how the past had turned into the present, the future would eventually turn desolate.

This is not to say that the idea of progress died with the birth of entropy. Some evolutionists, such as Herbert Spencer, continued to believe that progress, or development, is the grand law of the Universe. As evidence of the fact that scientific progress is not inevitable, however, we can offer the statistic that 46% of the population of the most scientifically advanced country in the world still believe the Genesis account of creation.

Since Darwin made his most famous statement about *De rerum natura* later than '1850', we hereby declare barnacles to be neither higher nor lower creatures, and nominate his four massive monographs on fossil and living barnacles (1851–54) as our books of the period. Being an evolutionist is not all fun and games.

New worlds in old (1900 – 14)

In 1887, Heinrich Hertz invented the first machine for generating and detecting radio waves. In 1913 Niels Bohr published the three-part paper that established the quantum theory of the atom. The years between saw the discoveries of X-rays, which opened the bodies of living people to medical scrutiny and gave scientists a powerful tool for investigating matter; α -rays, β -rays and radioactivity, which opened the heart of the atom to physicists and transformed the concept of element; the electron, which, as the



Eyes on the prize: Watson (far right) gets his share of the Nobel for discovering the structure of DNA.

fundamental entity of matter and electricity, brought together basic ideas of physics and chemistry; the elementary quantum of action, which gave the character to our jumpy century; the theory of relativity, which tied together space, time and gravity, and deprived all of them of their intuitive form and meaning; the internally structured atom; the particle theory of light; the tungsten filament, wireless telegraphy, isotopes and the all-electric kitchen. In a word, physical scientists entered the microworld, where they found more wonders than Columbus did in Cipangu.

The great discoveries in the microworld around the turn of the century seem to have followed one another as certainly as the trades blew the Admiral of the Ocean Sea to the West Indies. And, just like Columbus and his men, the first explorers of the microworld did not know where they had landed. X-rays, radioactivity, the electron, the quantum, and so on, seemed at first to have a place in the physics of the time. But by 1910 physicists had begun to realize that their novelties were profound, that they had not landed in Cipangu but in some place that could not be grasped with the concepts that had guided them there.

The trajectory of physics between Hertz and Bohr may be followed in the combined careers of J. J. Thomson and his student Ernest Rutherford, who together acted as Moses preparing the way for Bohr as Joshua. Thomson followed up the detection of X-rays by Wilhelm Röntgen to discover the electron. Rutherford began with an interest in radio waves, but when he entered Thomson's laboratory in Cambridge he was caught up in the investigation of X-rays and the rays from radioactive substances discovered by Henri Becquerel and the Curies. In trying to imagine how, on ordinary physical principles, the α particle interacted with common matter, Rutherford devised the nuclear model of the atom. Bohr began his postdoctoral work on the electron in Cambridge with Thomson, then switched to Rutherford at Manchester to learn something about radioactivity. While in Manchester, he adopted the nuclear atom and found a way, though tentative and crude, to use it as a guide to a new physics.

Not everyone agreed with Rutherford that science is either physics or stamp collecting, and the new biology was equally at home in the microworld. Between Theodor Boveri's demonstration of the individuality of chromosomes in 1888 and Thomas Hunt Morgan's identification of sex-linked inheritance in that Job of the modern laboratory, the lowly fruitfly *Drosophila*, the cell was positioned squarely at the centre of development, heredity, function and much else besides. Even before the rediscovery of Gregor Mendel in 1900, E. B. Wilson had pronounced the cell theory on a par with evolution as a biological generalization. Even after



Field work: choosing the site for CERN, Europe's entry into the Big Science league.

being rediscovered, Mendel had a rough ride, but by '1900' much had been done to clear the way towards a synthesis between the Mendelians and the biometricians, and thus the integration of experimental biology into the wider concerns of evolutionary theory.

If we must limit ourselves to a single representative compendium, it would be Rutherford's *Radio-Activity* (1904, 1905, 1912), as providing the basis for a radically new concept of matter. We make a gesture towards the other discoverers by adding the winners' lists of the Nobel institution, which gave its first awards in 1901. Most of the people mentioned under '1900' eligible for prizes under the statutes won at least one; Hertz, Joshua, Mendel and Moses, being dead, could not be considered.

Americanization (1950 – 7)

"With a big barrel of greenbacks." Thus the late Luis Alvarez — who designed the proton linac, built the first effective bubble chambers and X-rayed the pyramids — answered the question of how Berkeley led the postwar world in particle physics. In '1950' its accelerators, built with money from the US Army, the Atomic Energy Commission and the State of California, created the first mesons made by man (the muon and pion in 1947–48), the first pieces of antimatter (the anti-proton, 1955) and three Nobel prizes. Europe had to pool its greenbacks to compete. The result was the European Laboratory for Particle Physics, CERN, founded in 1954 under the leadership of the French and Italians. The British acceded to the inevitable with their customary good grace and enthusiasm for continental entanglements. Nonetheless, CERN is now regarded as an exemplar of European integration.

American physicists mined the spoils, or rather the surplus, of war in many ways.

Expertise and apparatus from the radar programme guided experiments that brought the maser and laser. Military interest supported the development of electronic computers. The Air Force supplied planes and electronics to explore the ionosphere and record cosmic radiation. National laboratories, with staffs of thousands and budgets of hundreds of millions, multiplied and grew: Argonne, Berkeley, Brookhaven, Livermore, Los Alamos, Oak Ridge. Physicists continued their wartime practice of working in groups of a size proportional to their machines; some grew to dozens and eventually to hundreds. The *Physical Review* became the most prestigious — and doubtless the largest — physics journal in the Western world.

As dinosaurs learnt, however, becoming larger is not necessarily the best way to survive, and by '1950' some physicists were already defecting to the 'softer' sciences of life. Max Delbrück, Erwin Schrödinger and Francis Crick all had a hand in the creation of molecular biology. The consequences of the paper of Crick and James Watson can be seen in the Genome Project and in the way that molecular biologists have begun to beat the physicists at their own game: by enlarging the scale of science, recruiting more scientists, raising more money, in a word, Americanizing.

The compendia of half a century ago are Henry De Wolf Smyth's *Atomic Energy for Military Purposes* (1945), which disclosed the scale of the Manhattan project to an astonished world; the pile of documents, scientific and political, drawn up to establish CERN; and that short 1953 communication, in this journal, which made the double helix the scientific icon of our age. ■

J. L. Heilbron is at Worcester College, Oxford OX1 2HB, UK. W. F. Bynum is at the Wellcome Institute for the History of Medicine, 183 Euston

The waning of the Golden Age

Politics, war and the decline of German science.

Einstein's German World

by Fritz Richard Stern

Princeton University Press: 1999. 271 pp.

\$24.95, £15.95

Walter Gratzer

"If my Theory of Relativity is found correct," Albert Einstein told his audience at the Sorbonne in 1920, "Germany will claim me as a German and the French will assert that I am a citizen of the world. If my theory proves false, France will say that I am German and Germany will declare that I am a Jew." Well, the theory, as we all know, was triumphantly vindicated by the solar eclipse expedition in 1919, and Einstein emerged as an international celebrity, the first citizen of Germany.

But many of his German *confrères* had no love for Einstein. In the first place he had denounced the patriotic frenzy for war that had gripped Germany in 1914; he had refused to sign the infamous manifesto of the 93 intellectuals, the so-called Fulda declaration, entitled *Aufruf an die Kulturwelt*, which affirmed the justice of Germany's cause and imperiously rejected accusations of atrocities by the German armies in Belgium. Einstein had even drafted a counter-manifesto, calling for an immediate end to the war, but could only muster three signatures besides his own.

And then there was relativity: Einstein's fame engendered a bilious resentment and, besides, the theory offended against the common sense (or *Anschaulichkeit*, as it was called) cherished by the classical physicists. The charge was led by two Nobel laureates, Philipp Lenard and Johannes Stark, both corroded by the anti-Semitism that permeated much of German society. At a meeting of the German Physical Society in 1920, Lenard launched a public attack on Einstein (whom he had at one time held in high esteem). Einstein unwisely replied in the tone of biting irony of which he was a master. Injurious men, it is said, brook no injuries, and Lenard's wounded self-esteem erupted in a froth of anti-Semitic paranoia. Relativity, he proclaimed, was (like quantum theory) a malign alien poison in the bloodstream of German science, a Jewish swindle.

Max von Laue, whose courage throughout the Nazi years was unflagging, told Einstein that when he expounded relativity theory in his student seminars, he always reassured the class that it had been originally written in Hebrew. Yet officialdom recognized Einstein's value as living proof of Germany's continued scientific ascendancy. "Funny people, these Germans," Einstein



Chaim Weizmann, the chemist who was destined to become the first president of the state of Israel.

noted in his diary. "I am a stinking flower for them and still they keep putting me in their buttonholes."

Such, then, is the background against which Fritz Stern's collection of linked essays is set. Of all the eminent historians of our time, Stern is best qualified to inform us about the German scientific scene, for he knew Einstein, and was a family friend and godson of Fritz Haber, a friend of Paul Ehrlich's widow and a cousin of Einstein's one-time assistant, Otto Stern (who later also won the Nobel prize). Fritz Stern's writing has an unmatched authority and a magisterial sweep that throws a brilliant light on the tragic disintegration of a noble culture, one in which science reigned supreme.

Stern sees the years between about 1880 and 1914 as a golden age, rivalling in its lustre the period known to the Germans as the Age of Genius. During that time, which Stern dates as about 1770 to 1830, German philosophy, literature, poetry and visual arts transformed the cultural landscape of Europe. The Wilhelmine golden age, by contrast, was marked by a flowering of technological innovation and scientific discovery. Many of the leaders were Jewish. Animosity against the Jews, by reason of what were perceived as unendearing social traits, was common, although seldom virulent; no more so, at all events, says Stern, than that of the Lutherans against the Catholics. (One of his chapters, in fact, is devoted to a trenchant refutation of the thesis propounded by David Goldhagen

in his book, *Hitler's Willing Executioners* (Little, Brown, 1996), that a murderous anti-Semitism was a universal fact of German culture and society.)

The vainglorious Kaiser Wilhelm II was, to be sure, fiercely anti-Semitic, but he recognized the contribution of Jews to science and to industry (and indeed ennobled his financier, the Jewish banker Gerson Bleichröder — the subject of Stern's superb study, *Gold and Iron* (Allen & Unwin, 1979)). For a man of such narcissistic self-absorption, bigotry and circumscribed intelligence, Wilhelm was remarkably receptive to the importance of science, both for itself and for what it might do to elevate German industrial and military power. He took an active part in founding the Kaiser-Wilhelm Society and its institutes (the precursors of today's Max-Planck Institutes) to promote science on a grand scale. Stern relates that, following his discovery of X-rays, Wilhelm Röntgen received from the Kaiser a telegram, which read: "I praise God for granting our German Fatherland this new triumph of science". God, fatherland and science, Stern comments, "the whole legitimating dogma in one sentence".

The Fulda declaration, crass as it was, was signed by such luminaries of science as Paul Ehrlich, Fritz Haber, Richard Willstätter (all Jews by birth) and the high-principled and pious Christian, Max Planck. Haber was, like so many of the assimilated Jews of the time, steeped in German culture and intensely

patriotic. (He was named indeed after Frederick the Great — *der alte Fritz*). Wholly godless, he converted to Christianity, in the interests of his career and to the consternation of his father. It profited him little; a friend offered the following explanation for Haber's failure to secure a professorial chair: "Before 35 he was too young, after 45 he was too old, and in between he was a Jew".

The chapter on Haber and Einstein is the centrepiece of Stern's book. Haber, cultivated, ambitious and cynical, had met Einstein, 11 years his junior, at a conference, and a deep friendship had developed between these two disparate personalities. When Einstein's first marriage broke up Haber became his confidant and go-between, and it was he who initiated the move to bring Einstein to Berlin. The closeness endured despite Haber's martial zeal when the Great War engulfed Europe. The war years were for Haber the most fulfilling of his life. He had originated the Haber-Bosch process for the fixation of nitrogen, without which, when the Atlantic blockade put an end to imports of 'Chile saltpetre' (sodium nitrate), Germany's war effort would have petered out in a matter of months.

Now Haber threw himself into the development of gas warfare, which he believed could bring a rapid and satisfactory end to the fighting. He took immense pride in his rank as a captain in the army, and exposed himself to much danger in the front line. His wife, an organic chemist, begged him to abandon his work on the hideous weapon that he had so eagerly espoused, and when he rejected her pleas she shot herself with her husband's service pistol. Haber returned next day to the front.

Haber was devastated by the collapse of Germany in 1918. There was even a move by the Allies to arraign him as a war criminal. Even Einstein was dismayed (like John Maynard Keynes) by the retribution the Allies sought to exact from his bankrupt and exhausted country. Haber, ever practical, devoted the resources of his institute to economic ends. These included an ill-starred project to extract gold from sea water in sufficient quantity to pay off the war debt, and the synthesis of insecticides, including, by a macabre irony of fate, the compound Zyklon B, with which the Nazis were to kill so many of Haber's friends and even relatives.

Haber is remembered now as one of the truly tragic figures of his time. Einstein did not wait for Hitler's accession to leave Germany, but Haber stayed until conditions became intolerable in 1933. Einstein wrote him a letter of sympathy, but was unable to resist adding a barbed remark or two: he was glad that Haber's "love for the blond beast" had cooled a little, and he rather callously likened his friend's predicament to giving up a theory on which one had worked all one's life.



Einstein (right) and Haber: a deep friendship developed between these two disparate characters.

Chaim Weizmann, the chemist who was destined to become the first president of the state of Israel, had held Haber in low regard, for he had indeed not concealed his distaste for the Zionist cause and had urged Einstein not to support it either. But the sick and exiled Haber was a chastened man. As Weizmann tells it in his autobiography, *Trial and Error* (Greenwood Press, 1972), Haber confessed to him: "I was one of the mightiest men in Germany. I was more than a great army commander, more than a captain of industry ... All doors were open to me ... At the end of my life I find myself a bankrupt. When I am gone and forgotten your work will stand, a shining monument in the history of our people." Haber accepted the offer of shelter in the Daniel Sieff Institute (later to bear Weizmann's name) in Palestine, but a few months later he was dead. He was at least honoured in a memorial service, organized by Max Planck — who at other times prevaricated and compromised with the regime for the sake, as he saw it, of German physics — in defiance of instructions from the ministry. Such resolve was never displayed again.

Weizmann's political struggles form the subject of another of Stern's chapters. His work in England during the Great War led to a fermentation process for the production of acetone, an intermediate in the manufacture of explosives. This brought him recognition and the friendship of Winston Churchill and Arthur Balfour, from which flowed the Balfour Declaration and in due time the state of Israel.

Einstein initially embraced Weizmann's aims, but later began to have doubts. In

particular, he was concerned for the fate of the Arabs in Palestine, and he warned Weizmann against the growth of a militant Jewish nationalism — *à la prussienne*, as he termed it. "If," he continued, "we do not find the path to honest cooperation and honest negotiation with the Arabs, then we have learned nothing from our 2,000 years of suffering, and we deserve the fate that will befall us." Prescient words. Weizmann, goaded beyond endurance by Einstein's continuing objurgations, called him "a prima donna who has lost her voice" and severed their association. It was Weizmann's tragedy to lose the argument with those far less liberal and accommodating to the Arabs than himself, and, worse, to be betrayed by the British, who obdurately refused Jewish refugees a haven in Palestine.

Stern expatiates also on the career of another tormented German Jewish intellectual, the industrialist and foreign minister of the Weimar Republic, Walther Rathenau, who was assassinated in 1922 by fascist thugs. He concludes his journey through twentieth-century Germany with two moving essays. The first reflects on how historians who fought or were otherwise enmeshed in the Great War reacted to their experiences, the second on the tragic displacement of populations, especially Germans and Poles, after the Second World War.

Stern's is the civilized voice of reason and understanding; his book is revealing, absorbing and often poignant. Scientists will surely be captivated, as I was, by Stern's accounts of four of his scientific heroes — Einstein, Haber, Ehrlich and Planck. All were in their several ways tormented spirits. They suffered personal tragedies and adversity. All found refuge of a sort in their science, but that too exacted a heavy price. Friendship and collegiality proved more important than their families, to whom none but Planck gave much time or love. Einstein wrote to his cousin Elsa, when his first marriage was failing, that his love for science flourished "because it lifts me out of my vale of tears to quiet spheres, impersonal and without rebuke or complaint". All these men drove themselves to the point of collapse. They would, Stern avers, have understood Sigmund Freud's dictum that "science betokens the most complete renunciation of the pleasure-principle of which our minds are capable". Science as a vocation, Stern reminds us, is far removed from science as a career.

Times change. Look around you in the coffee-room and ask yourself how many of your colleagues adhere to such austere virtues. For my part, the vignette of Einstein most pleasing to recall comes from the memoirs of Arnold Toynbee (*Acquaintances*, Oxford University Press, 1967). While on a visit to Oxford, Einstein was observed by the professor of Greek, Gilbert Murray, sitting in

a quadrangle of Christ Church college “with a far-away look on his face. The far-away thought behind that far-away look was evidently a happy one, for at that moment the exile’s countenance was serene and smiling. ‘Dr Einstein, do tell me what you are thinking,’ Murray said.

“‘I am thinking,’ Einstein answered, ‘that after all this is a very small star.’” ■

Walter Gratzer is at the Randall Institute, King’s College, University of London, 26–29 Drury Lane, London WC2B 5RL, UK.

Beneficent squid

The Squid Synapse: A Model for Chemical Transmission

by Rodolfo R. Llinas

Oxford University Press: 1999. 224 pp.

£49.95, \$85

Tomoyuki Takahashi

When it comes to contributing to our knowledge of neuronal function, no creature can surpass the squid. Its axon has enabled us to discover the ionic basis of electrical signalling, and its synapse has revealed the chemical signalling between neurons. Rodolfo Llinas’ book summarizes what this invertebrate has told us about chemical synaptic transmission.

The author is a leading scientist in research fields ranging from cellular to system neurophysiology, and the first half of his book introduces readers to classic electrophysiological findings on chemical transmission, many of which cannot be found in general textbooks. The rest of the book is devoted to the molecular mechanisms of synaptic transmission, in particular the roles of synapsins and SNAREs in transmitter release.

Throughout, much emphasis is given to the key role of calcium, which is described with many original findings by the author and his collaborators. In this respect, the book may be regarded as the notebook of a scientist who has pursued the “painstaking deciphering of nature’s small print”, as the author himself puts it. Despite the book’s personal nature, it successfully reveals the current state of synaptic research and the direct approaches that have been used to build this up. It should therefore attract a wide audience, particularly those interested in listening to nature directly rather than passively accepting current paradigms.

Finally, the book comes with a CD-Rom containing an excellent program of a synaptic model. The program enables one to perform ‘virtual’ recordings from a presynaptic terminal and postsynaptic target cell, and to test the effects of manipulating various parameters such as intracellular and extra-



Axon at work: the squid’s oversized neuron has contributed much to neurophysiology.

cellular calcium concentrations or the shape and size of presynaptic action potentials. This program will surely attract young students, as well as researchers, who are interested in computer simulations of neuronal function. This model also reminds us of the dynamic aspects of synaptic function and that individual neurons are not merely passive elements in neuronal networks, but actively influence brain function through synaptic modulation. ■

Tomoyuki Takahashi is in the Department of Neurophysiology, Faculty of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

Medicine’s least respectable branch?

A Century of Psychiatry, Volumes I and II

edited by Hugh Freeman

Mosby-Wolfe: 1999. 360 pp. \$120, £59.95

Dylan Evans

The history of psychiatry is a complex and fascinating affair, but it has not been easy for the general reader to obtain an accurate picture of it. Most accounts have been either accurate and academic, or popular and tendentious. With *A Century of Psychiatry*, the general reader now has a reference source that is both accessible and balanced.

This multi-author, twin-volume work does not shrink from the darker moments in the history of psychiatry, from the reckless enthusiasm for poorly tested remedies in the

1930s and 1940s, to the political abuse of psychiatry in the former Soviet Union. However, these are not the basis for any sweeping pronouncements about the nefarious purposes of psychiatry à la Michel Foucault. Rather, the constant references to the wider social and scientific context in which these events occurred helps to make them understandable, even if not always excusable.

In his excellent chapter on the origins of convulsive therapy, for example, Max Fink makes palpable the sense of impotence felt by psychiatrists in the first decades of this century, when no effective treatment for the psychoses was known. Fink also explains the reasoning that led doctors to the idea of treating schizophrenia with artificially induced fits: they were guided by the theory of the antagonism of diseases, and thought that schizophrenia was antagonistic to epilepsy (an idea that also led to the converse therapy, namely the attempt to halt epilepsy with transfusions of the blood of psychotic patients). This is a welcome corrective to the misleading images of convulsive therapy perpetuated by popular culture.

The balanced approach is evident throughout the book. The sections on the emergence of psychoanalysis and Sigmund Freud by Martin Stanton, for example, avoid the extremes of hagiography and demonography that typically afflict works on these subjects. Anyone who baulks at the idea of including Freud in a history of psychiatry should read Gerald Grob’s section on American psychiatry in the 1960s, which makes clear the extent to which psychoanalytic ideas dominated the treatment of mental disorder there for several decades. As Charles Cahn’s section on the history of

the *Diagnostic and Statistical Manual of Mental Disorders (DSM)* shows, it was not until the publication of *DSM-III* in 1980 that most psychoanalytic concepts were formally excluded from psychiatric orthodoxy in America. *DSM-III* replaced the old psychoanalytic concept of 'neurosis', which had been central to *DSM-I* and *DSM-II*, with the more general one of 'disorder' (although the word 'neurosis' was restored in parentheses after protests by psychoanalysts fighting a rearguard action).

The organization of the text is partly chronological and partly thematic. There are 10 chapters, each dedicated to one decade of the twentieth century. Each chapter consists of various sections which cover particular themes, such as a form of treatment, a diagnostic category, a famous psychiatrist, or some wider social or national development. Each section can be read on its own, making the book a handy reference tool. There is inevitably a degree of overlap between the various sections, but this is not a serious defect. The book is amply illustrated with archive photographs, which bring the stories vividly to life. There are no references in the text, as befits a non-academic work, but there are plenty of suggestions for further reading at the end of each chapter.

Perhaps the most interesting thing to emerge from this collection of historical essays is a real sense of the nitty-gritty of scientific discovery. The mental disorders are such terribly complex phenomena that they are still poorly understood today. As this book shows, there has been real progress in understanding and treating mental illness during the past century, but it has been a very haphazard advance. Many therapies that initially seemed promising turned out to be dead-ends, while the few therapies that have stood the test of time, such as neuroleptics and antidepressants, were largely discovered by accident.

Intuition and enthusiasm have played as important a role as theory in driving psychiatry forward, and the history of the discipline is correspondingly peppered with colourful characters and leaps of imagination. The fact that psychiatry has had more than its fair share of eccentric geniuses may be one of the reasons for its failure to shake off its image as the least respectable branch of medicine, but what progress there has been is almost entirely due to these mavericks. There is a view of scientific progress that downplays the context of discovery. The history of psychiatry favours a different view, one that sees nature's mysteries as giving way to an odd combination of brilliant mental leaps, accident and sheer, dogged persistence. ■

Dylan Evans is in the Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK.

In retrospect: chosen by David Jones

The Year 2000: A Framework for Speculation on the Next Thirty-three Years

by Herman Kahn and Anthony Wiener
Macmillan: 1967. £3.15

The above volume was perhaps the most professional attempt made during the 1960s to predict the world of AD 2000. Its authors worked at the Hudson Institute, a respected US independent think-tank, and their remit covered several aspects of the world 33 years ahead.

One aspect was economic — how rich the world, and the different countries within it, would be by 2000. Knowing that the economy, and society in general, was driven largely by scientific and technical developments, the tank-thinkers made guesses in these areas too. They were greatly concerned by politics and wars. They even tried to predict sociological changes — the feel and character of society in the future. Like all sensible crystal-ball gazers, they hedged their bets. They expounded not only the most likely 'surprise-free projection', but a range of spreads around it, together with many possible surprises and variants. Their reference region was, of course, the United States, but their scope was global. How well did they do?

In economics, their forecasts tended to be on the high side. Thus, for the United States, they predicted a GNP for the year 2000 of (in 1965 dollars) \$1.3–4.5 trillion, with a best guess of \$3.2 trillion. Taking \$1 (1965) = \$5.14 (1998), this is \$7.1–23 trillion, with a best guess of \$17 trillion. The published figure for 1998 was \$8.5 trillion.

Their view of the technological future was similarly bold and expansive. They listed 100 significant developments they expected before the year 2000, and feared they would miss many others. In fact, fewer than 30 of their 100 have come to pass, mainly without much impact. They also suggested 35 "less likely" possibilities, such as direct input into the human memory, a drug equivalent to Aldous Huxley's "soma", and genetic modifications taking human beings beyond the definition of *Homo sapiens*. Perhaps mercifully, not one of their less-likely list has yet been realized.

Kahn and Wiener's few accurate predictions dealt mostly with electronic communications. They successfully foresaw home video-recording, automated banking systems, pervasive high-capacity data transmission, mobile phones, satellite broadcasting, the universal business use of computers, and new pervasive techniques for the surveillance, monitoring and control of individuals and organizations. But they did not miss as much as they had feared. The most noticeable absentees are the global positioning system, the personal computer (instead they envisaged universal 'thin client' terminals, served by central time-sharing mainframes), the pocket calculator and the video game. Technology has

not advanced as dramatically as they had guessed.

In their political prophesies, the tank-thinkers concentrated on the international scene — war and inter-bloc conflicts. They projected the cold war forward from 1967 in many ways, some culminating in nuclear exchanges (a professional concern of the Hudson Institute at the time), and others maintaining a tense stability. Some of their scenarios featured an erosion of communism, but none its outright collapse. Only the simple-minded genius of Ronald Reagan ever imagined that outcome.

But the most interesting speculations in the book are sociological. Here the dangerous optimism of the 1960s shines out clearly. Kahn and Wiener's predictions of economic growth, and the boundless wealth to come, led them to predict a "post-industrial society" whose main problem would be making the best use of leisure time. As they grew richer, such societies would evolve into a sort of hippy paradise, with perhaps half the population more or less voluntarily and cheerfully loafing on the vast output of an automated industrial system. Thirty-three years later, that boundless wealth, in GNP terms at least, has largely come our way; and yet we have somehow failed to enjoy or even notice it. Most working people feel as pressured as ever, perhaps more so. And the increase of the underclass, never anticipated by the most doom-laden prophets, has disfigured even the richest modern societies.

I cannot blame the tank-thinkers for failing to foresee the ramifications of the social trends of their day. The 1960s hippies, pursuing happiness as the Declaration permits them, caught it up in the form of addictive drugs — and lit a vast social conflagration that is still spreading. The feminist revolution, the environmental movement, the growth of global migration and ethnic cleansing, the eruption of AIDS: all these would have been beyond the most insightful of prophets.

Kahn and Wiener did, however, glimpse the spread of electronic totalitarianism. Their technical predictions included some shrewd guesses at how, by the merging of big databanks and the widespread installation of surveillance equipment, whole populations could be routinely and continually monitored. They clearly foresaw the erosion of privacy. To them this was an evil, a threat to freedom. Our more enlightened age knows it as a safeguard. Only the gaze of innumerable cameras saves us from the underclass. Only by allowing the government to peruse all our bank accounts can we discourage the tax-dodgers and the money-launderers. Only when every telephone conversation and every Internet connection is logged, and even recorded, can the gangsters, terrorists and pushers be kept in check. Freedom must sometimes be destroyed to save it. ■

David Jones is in the Department of Chemistry, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, UK.

Making waves in physics

Three wave singularities from the miraculous 1830s.

Michael Berry

Singularities are places where mathematical quantities become infinite, or change abruptly. In waves — of light, for example — there can be singularities in the intensity, in the phase, or in the polarization. This is a modern view, sharply different from the traditional approach where waves were simply the solutions of wave equations, and singularities — if considered at all — were regarded as awkward places where the usual treatments fail. And yet the foundations of wave singularities were laid in three astonishing papers published as early as the 1830s, a decade whose intellectual significance we are only now beginning to appreciate.

Reversing historical order, we start in 1838, with a paper by George Biddell Airy. The immediate stimulus was the theory of the rainbow. Two centuries earlier, Descartes had understood the bright bow as Sun rays directionally focused by raindrops. This geometrical theory gives a good first approximation but fails to account for the delicate supernumerary bows sometimes seen just inside the main arc. In 1801 Thomas Young realized that by regarding light as waves it is possible to understand supernumeraries as interference fringes, but could not give a precise mathematical description. Airy's contributions were: to appreciate that the rainbow is a particular example of a caustic, that is, a line where light rays are focused (the bright lines on the bottom of swimming pools are also caustics); to see that caustics are singularities, where ray optics predicts infinite brightness; to realize that wave physics will soften the singularities; and to discover the precise mathematical description of this softening, in the form of his rainbow integral (pictured above).

Airy's paper was doubly influential. First, because refined techniques soon devised by George Gabriel Stokes to study the rainbow integral established divergent infinite series as an important tool in bridging gaps between physical theories (in this case ray and wave optics) and uncovered a mathematical phenomenon whose ramifications are still being explored. Second, because the rainbow integral, describing wave interference decorating the simplest kind of caustic, was found to be the first in a hierarchy of 'diffraction catastrophes'. These more elaborate wave singularities are now classified using catastrophe theory — mathematics whose application greatly advances the physics of caustics.

Next are two papers from 1833 and 1836 by the polymath William Whewell (to whom



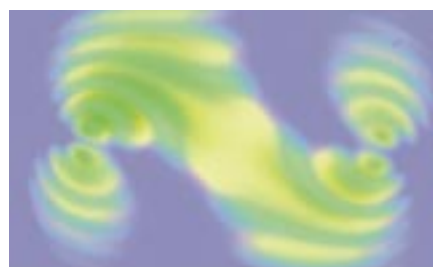
Airy's rainbow integral.



Whewell's amphidromy between England and Holland.

we owe the word 'physicist', with its "four sibilant consonants that fizz like a squib"). He was studying the tides in the oceans, and seeking to "connect the actual tides of all the different parts of the world — and to account for their varieties and seeming anomalies". From Young he learned to concentrate on the cotidal lines connecting places where the tide is high at a given time. Cotidal lines are wavefronts of the tide, regarded as a wave of twelve-hour period; they are contours of the phase of this wave. Whewell appreciated that a map of cotidal lines would render intelligible the pattern of tides around the coasts, and coordinated hundreds of new observations, in an early international scientific collaboration.

Extrapolating away from the coasts into the 'German Ocean' he reached the extraordinary conclusion that there must be "rotatory systems of tide-waves [where] the cotidal lines ... revolve around [a point] where there is no tide, for it is clear that at a point where all the cotidal lines meet, it is high-water equally at all times". One of Whewell's 'amphidromies' can be seen above. What Whewell discovered were the phase singularities of the tide waves.



Hamilton's diabolical points (bullseyes) in several square centimetres of overhead-projector transparency foil viewed obliquely through crossed polarizers; in each bullseye, the interference rings are contours of difference of wave speeds, centred on an optic axis, and the black stripes reflect geometric phases.

Phase singularities are now recognized as important features of all waves; in three dimensions they are lines rather than points. At a phase singularity, the wave intensity is zero — in contrast to caustics where the intensity is (geometrically) infinite. In acoustics, the singularities are threads of silence; in light, they are optical vortices; in superfluids, quantized vortices; and in superconductors, quantized lines of magnetic flux.

In addition to intensity and phase, waves of light exhibit polarization: they are vector waves. Polarization has its singularities too, discovered by our third author, William Rowan Hamilton, in 1832, as an unexpected consequence of Augustin Fresnel's theory of the optics of crystals. In a general anisotropic material, two waves propagate in each direction, with different speeds and polarizations. There are, however, two singular directions, optic axes, where the speeds are the same. As functions of direction, the two speeds can be represented by surfaces forming a diaboloid (double cone) at each optic axis. Hamilton deduced that at such 'diabolical points' the wave direction corresponds to a cone of rays. This unprecedented 'conical refraction' was soon observed, confirming that light is a transverse wave. Diabolical points can easily be seen directly (pictured above).

Hamilton's diabolical point was the first physical example of degeneracy between eigenvalues of a real symmetric matrix. Its descendants thrive in many areas of science today, for example, as conical intersections between energy levels in quantum chemistry, as Bloch wave degeneracies in the quantum Hall effect, and as the simplest geometric phases.

Michael Berry is in the Department of Physics, University of Bristol, Bristol BS8 1TL, UK.

How science saved the world

Has science driven history for the past 50,000 years?

Kim Stanley Robinson

Review:

Science in the Third Millennium
by J. S. Khaldun

Scientific Careers 2001–3000
by AI 80004 “Ferdnand”

The 12 preceding volumes of Professor Khaldun’s study articulate his belief that science has been the driving force in human history for the past 50 millennia at least, an idea sometimes criticized as containing too broad a definition of science to be useful. His latest book, however, brings him into an era when the idea can be tested for its explanatory power.

It’s also a period that illustrates Khaldun’s contention that science was always a politics by other means, a utopian and unselfaware revolution attempting to shift power from residual aristocracies to the general populace. What early scientists took to be neutral and necessary aspects of the scientific method, such as reproducibility, falsifiability, Occam’s razor, peer review, and so on (cf. Khaldun pp. 580–935), were actually political institutions all along, attempting to regulate human action with the ultimate goal of alleviating suffering. The invention of science as a method for understanding nature obscured for centuries the realization that it was also the optimum mode of social regulation. Only the population-overshoot crisis early in the millennium brought scientists to an awareness of this.

The overshoot, as Khaldun reminds us, was a dangerous time. Humanity found itself in a race to invent a sustainable way of life

before its reproductive success and primitive technology severely damaged the carrying capacity of the planet. Global warming and an anthropogenic major extinction event did occur, but it could have been worse: a full six billion of our ancestors existed simultaneously at the tip of an unsustainable technological prosthesis, and the social system facing this situation was the Late Feudal phase known as capitalism. Many technical advances of the era therefore did nothing to mitigate the ecological disaster or, worse, added to it. Even the great leaps in the biological sciences, reinforcing health and initiating the assault on senescence, only exacerbated matters, as humanity entered the age of longevity still enmeshed in a dysfunctional social system.

In describing this crux, Khaldun rehearses again his theory of history in which social systems consist of clashes between residual and emergent elements. He reads feudalism as a clash of traditional command economies with emergent capitalism; capitalism as a clash of feudalism with early permaculture; and our late permaculture as a clash between residual capitalism and some poorly modelled emergent harmony which for biological reasons may be approachable only asymptotically.

In the overshoot crisis, Khaldun sees this struggle expressed as a conflict between science, representing the emergent permaculture, and capitalism, representing residual elements of feudalism.

Capitalism attempted to maintain a hierarchy in which science would serve as pet monkey, cranking out new commodities and increasing lifespans. Science resisted this

impulse not only because of the practical danger of the overshoot to the progeny of scientists, but also because science itself would be threatened if the residual elements succeeded. So scientists of the era, despite the lack of a paradigm describing them as historical actors, set to work transforming capitalism into a system of practices they recognized as more rational, universal, lawful and pragmatic — that is to say, more scientific.

The resulting ecological reorganization of the economy and increasing advisory power of science in the political realm initiated the institutional mutation towards the system identified retroactively as permaculture, which now seems to us so natural. Adequate food, water, shelter, clothing, health care, and education for all; increased longevity; the stabilization of the Earth’s ecologies; the inhabitation of the Solar System: these and other major achievements of the millennium all followed from science’s successful transformation of the social system of that time.

Whether macrohistories such as Khaldun’s can still convince is doubtful. His “metameme narrative” of residual and emergent smacks of Hegelian teleology, in which progress happens no matter what people choose to do. A useful corrective to this kind of metaphysics is the recent “hyperanalyst” work of the computer archaeologist AI 80004 “Ferdnand”. Excavations in the enormous mass of data remaining from the overshoot period enabled 80004 to list all scientists working at that time. Their careers are summarized, then analysed in a series of tables and graphs charting their profits and investments, their political actions, their child-rearing, the ecological utility of the scientific content of their careers, and so on (cf. pp. 7378–9417).

Close perusal of these tables will indeed reveal which scientists contributed to the survival of the overshoot and invention of permaculture, and which did not. But one sees in the tables no teleology, no residual or emergent elements, no grand battle of metamemes — only small individual actions, spread across the whole range of possibility. Khaldun might claim that these actions “summed over history” reveal the beliefs motivating scientists, and thus the powerful existence of the metamemes. But this would be a subject for yet another volume, called perhaps *History Judges You*. ■

Kim Stanley Robinson is the author of Red Mars, Green Mars, Blue Mars and Antarctica. His latest book is The Martians.

e-mail: ksr@davis.com



Optics adapt to the whole sky

Brent Ellerbroek and François Rigaut

Atmospheric distortion limits the sensitivity of ground-based telescopes. Adaptive optics is the answer — new techniques mean that computer-controlled mirrors may soon compensate for distortion in any part of the sky.

Over the past decade, astronomers have made significant progress in improving the resolving power of ground-based optical and near-infrared telescopes. Atmospheric turbulence — more precisely, refractive index variations due to local fluctuations in atmospheric temperature and density — distorts light waves received from distant objects and severely limits the image quality. For large 8- to 10-metre telescopes in good locations, uncompensated atmospheric turbulence typically makes the resolution 50 times worse than could be seen by the same telescope in space. Fortunately, these optical distortions can now be corrected in real time by adjusting the surface of a thin flexible mirror. The commands to this mirror are determined by measuring the distortions in light waves received from a bright 'guide' star. This technique, known as adaptive optics, has so far been limited to regions of the sky where there are bright stars. A paper by Ragazzoni *et al.*¹ on page 54 of this issue shows that this technique may soon be applied to the whole sky.

Conventional adaptive optics is now used by a number of large ground-based telescopes, and is pushing resolution and signal-to-noise ratio towards what is possible in space. In June last year, the 8-metre Gemini North infrared telescope successfully separated images of Pluto and its moon Charon, using Pluto itself as the guide star for the adaptive optics². This is equivalent to resolving the separation between a set of automobile headlights from 2,000 miles away. But standard adaptive optics has several drawbacks. The guide star must be bright and very close to the object to be imaged. Vertical variations in turbulence mean that light waves from different directions experience different distortions, and mirror commands computed for a guide star in one part of the sky will only compensate turbulence within a very small field-of-view. In fact, only 0.1 to 1% of the sky is close enough to a bright star to benefit from adaptive optics³. Even within the immediate neighbourhood of a guide star the performance of adaptive optics is not always uniform, producing images with spatially varying resolution that can be difficult to interpret.

Astronomers are beginning to improve adaptive optical systems with artificial guide stars to overcome the limited sky coverage by naturally bright stars. Laser-generated guide

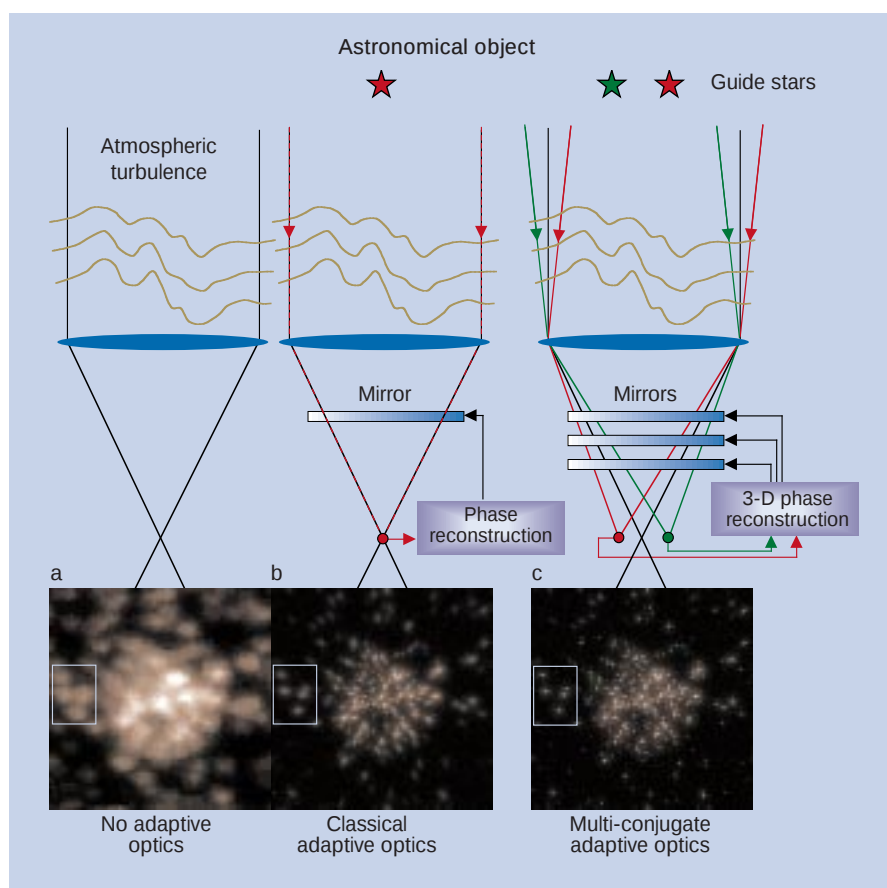


Figure 1 Adaptive optics broaden the field. a, Atmospheric turbulence blurs astronomical images from ground-based telescopes. b, Adaptive optics, by means of a deformable mirror and an optical-distortion sensor, compensates for the image degradation in a small field around the guide star. c, Using several deformable mirrors and several guide stars to compensate for turbulence in a three-dimensional fashion, multi-conjugate adaptive optics extends the image compensation to a larger field-of-view.

stars can be created 90 km above a telescope, using a laser tuned to the sodium D_2 line to excite sodium atoms in the atmosphere^{4,5}. With a sufficiently powerful laser the photons emitted as the atoms return to their ground state can serve as a guide star. Such systems face big technical challenges: from developing powerful and reliable lasers and projecting the laser beam to the sky with minimum distortion, to accounting for temporal variations in the distribution of the sodium layer. Despite these complications, several adaptive optical systems equipped with fairly modest lasers have already achieved results within a factor of two of their theoretical limits⁶.

The use of lasers, however, does not

improve the uniformity of adaptive optical performance in the neighbourhood of the guide star. A concept known as multi-conjugate adaptive optics (MCAO) was proposed as early as 1989 to address this problem⁷. With current systems, a single mirror is deformed to compensate for distortions in the light from a single guide star. In MCAO, multiple mirrors are able to correct for atmospheric turbulence in three dimensions, in principle producing uniform compensation over a much larger field-of-view (Fig. 1). To calculate the three-dimensional distribution of the turbulence, and to command each mirror correctly, requires optical distortion measurements from multiple guide stars (either natural or artificial), each



100 YEARS AGO

The *Daily Chronicle* recalls that a London paper of the first week of 1800 alluded to the then recent hot disputes in France and England respecting the beginning of the nineteenth century. According to the paragraph, the famous Joseph Jérôme Lefrançois de Lalande, who then occupied the Chair of Astronomy in the University of Paris, had taken an active part in the controversy, and he had pronounced in favour of January 1, 1801. His decision had been generally accepted as correct on both sides of the Channel. The newspaper referred to remarks: "The same ridiculous question was agitated in 1700." So does history repeat itself.

From *Nature* 4 January 1900.

50 YEARS AGO

For some ten years, at the Zi Ka Wei Observatory, in Shanghai, China, we have been using, with exceptional success, a new technique in our weather forecasting. This method is based on an as yet unexplained correlation between the usual ionosphere echoes (E , F and F_2) and the future movement or behaviour of the three main air masses, which make the weather all over the world: polar, maritime and tropical, or equatorial, as some people call it. Our results have already been published in the *Bulletin of the Meteorological Society of America*, in 1946, and in a paper sent to the last Pacific Science Congress held in New Zealand last February. We found, in 1939, in Zi Ka Wei that by pulsing on the mean critical frequency for the E -layer at our station (6 Mc.), any day of the year, after sunrise and before sunset, we could forecast what type of air mass would make the weather over an area of 400 km. around Shanghai. For example, we found that when we had an E -echo, the maritime (Pacific Ocean) air mass would either stay over us, if it was already there, or it would come over us, if at the time of pulsing another type of air mass was covering our regions. When we obtained an F -echo, the polar (Siberian) air mass would make the weather, while an F_2 -echo meant the arrival of the tropical air, or its staying if already there. Knowing by this technique the type of air mass due to make the weather, we could forecast its future characteristics, dry or damp, overcast or clear, windy or calm, good or bad visibility, high or low temperatures, etc.

From *Nature* 7 January 1950.

of which provides a measurement of the turbulence in a different direction. The computer-intensive task of converting these linear projections into the original three-dimensional profile is known as atmospheric tomography.

Ragazzoni *et al.*¹ have shown that the tomographic method works. They collected optical-distortion measurements from four stars in the constellation Aquila, and used least-squares techniques to predict atmospheric distortions for the central star as a linear combination of the other three measurements. The mean-square error in this prediction is 17 times smaller than the initial distortion, and three times smaller than the error obtained when the least-squares method is replaced by an arithmetic average. Two deformable mirrors could use these predictions to correct for distortions in real time for the entire constellation. The inferior results obtained by averaging are comparable to what could be obtained with a single mirror and guide star.

Studies to assess the feasibility of MCAO for existing telescopes are already underway. For an 8-metre telescope, a system with three mirrors and five laser guide stars would mean a ten-fold increase in the area of the compensated field-of-view compared with conventional adaptive optics, and would dramatically improve the uniformity of correction within this area. The individual mirrors and optical-distortion sensors required to set up such a system are already available. Existing lasers are not yet up to the mark, but are being developed in the right direction. Calculations in real time would need to increase up by a factor of 15 to make atmospheric tomography possible, but processing on this scale has been demonstrated by other

adaptive optics programs. The use of MCAO for solar astronomy is also under investigation. In this case, guide stars are not needed, because optical distortions can be determined from images of the granulation in the solar photosphere⁸.

In the longer term, MCAO may be an important technology for the much larger 30- to 100-metre ground-based telescopes currently under discussion⁹. How the number of guide stars and deformable mirrors will scale with telescope size is not well understood, but it may be that the maximum angular separation allowed between guide stars for effective tomography is proportional to telescope diameter. If so, the relatively low angular density of bright natural stars would suffice for atmospheric tomography with 100-metre telescopes, eliminating the need for lasers. Of course, the technical advances needed to develop deformable mirrors and optical-distortion sensors for telescopes of this size are considerable. Getting to grips with the possibilities and requirements of such systems will be an important task for the adaptive optics community in the years ahead.

Brent Ellerbroek and François Rigaut are at the Gemini Observatory, 670 North A'ohoku Place, Hilo, Hawaii 96720, USA.

e-mail: bellerbroek@gemini.edu

1. Ragazzoni, R., Marchetti, E. & Valente, G. *Nature* **403**, 54–56 (2000).
2. Gemini North <http://www.gemini.edu/dedication/dedication.html>
3. Rigaut, F. & Gendron, E. *Astron. Astrophys.* **261**, 677–684 (1992).
4. Foy, R. & Labeyrie, A. *Astron. Astrophys.* **152**, L29–L31 (1985).
5. Happer, W., MacDonald, G. J., Max, C. E. & Dyson, F. J. *J. Opt. Soc. Am. A* **11**, 263–276 (1994).
6. Davies, R. *et al.* *Exp. Astron.* (in the press).
7. Beckers, J. M. *Proc. SPIE* **1114**, 215–217 (1989).
8. Rimmele, T., Radick, R., Dunn, R. & Richards, K. *NOAO Newsletter* **57** (1999).
9. Gilmozzi, R. *et al.* *Proc. SPIE* **3352**, 778–791 (1998).

Conservation biology

Living on the edge

Thomas Brooks

In an undisturbed ecosystem, populations at the edges of species' ranges are less dense and more variable in space and time than at their centres. So they are more prone to extinction¹. But surprisingly, accumulating evidence suggests that edges are where the last remaining populations of many species survive^{2,3}. On page 84 of this issue, Channell and Lomolino⁴ extend their previous work on this apparent paradox. For 245 species across a wide variety of taxonomic groups and geographical regions, they show that the collapse of ranges towards the periphery is the rule, not the exception. Their explanation is that human impacts, such as deforestation, moving in a front across the landscape, simply overwhelm the underlying

geographical ecology. Many species have life on the edge thrust upon them.

Channell and Lomolino chose 245 species for which maps of current and historical ranges are available. For each they located the centre point in a species' historical range as that most distant from any range edge. They defined each species' range periphery as everywhere more than half this distance away from its centre. They then calculated a species' 'index of centrality' as the proportion of its recent range that falls within the central portion of its historical range. Two-thirds of their species survived preferentially in their historical peripheries, and one-third only in these peripheries. The pattern of survival on the edge is a general one.

An example is the black-footed ferret (*Mustela nigripes*, see Fig. 1 on page 84), which now survives only in the northwest corner of its historical range.

How much trust can we put in the estimates of historical ranges? Mapping the contemporary ranges of species is hard⁵. So too is documenting past extinctions, as the work of the Committee on Recently Extinct Organisms⁶ exemplifies. Assessing historical distributions combines these two difficulties. Historical ranges will tend to be less finely resolved, and will approximate the extent of a species' occurrence rather than its area of occupancy⁵. They will therefore appear less fragmented than current ranges. (This problem is scale dependent, because it is easier to map the ecologically marginal areas of a species' range at fine scales.) This should not alter Channell and Lomolino's⁴ indices of centrality, however, because the effect of increasing the proportion of historical periphery should be balanced by the effect of the increasing proportion of current range within this periphery.

There are two places that show no tendency for peripheral populations to survive. The first, Hawaii, illustrates a case of range implosion, where threats simultaneously affected all of the coastal edges of species' ranges, pushing their last populations into the central uplands. In fact, Hawaii may not even be an exception, if considered from the perspective of ecological occupancy as well as geographical extent. Olson and James⁷ have suggested that before humans arrived many Hawaiian species were actually confined to the lowlands, and so had doughnut-shaped ranges. To the last remnants of Hawaii's animals and plants, the uplands may be the ecological periphery of their ranges, even if they are the geographical centre.

Africa is the other exception where species show no significant tendency to survive at the edges of their ranges. Channell and Lomolino's explanation is that people have a longer history in Africa than anywhere else⁸. Thus, human activity is already distributed throughout the landscape, instead of moving across it by contagion. This cannot be the whole explanation, however, because the declines of all of Channell and Lomolino's continental species are recent, rather than stemming from first human contact. (By contrast, the Pacific islands in their study were reached by people only in the last two millennia⁷.)

An alternative possibility is suggested by the taxonomic composition of the data. Nearly all (20 out of 24) of the African species are large mammals, compared with less than half in other regions⁴. Human hunters prefer large species⁸, and so have caused these range collapses in Africa. Elsewhere, the examples of extinction documented by Channell and Lomolino are usually the result of habitat destruction and



Figure 1 The extent of current and historical habitat in three 'hotspots'¹⁰. Surviving habitat tends to lie in the historical periphery. a, The Serra do Mar and the Paraná basin of the South American Atlantic forests (red line, historical range; green, surviving habitat). b, The northern Luzon and southern Mindanao in the Philippines (light green, historical range; dark green, surviving habitat). c, The extreme west and east of the Guinea region of West Africa (red line, historical range; green, surviving habitat).

invasive species. Could overhunting drive populations inwards, whereas other threats move laterally across the landscape? This suggestion is eminently testable by comparing the pattern of range collapse with body size of the species concerned or (more directly) with the human threat at work in the landscape.

Channell and Lomolino's work⁴ is obviously relevant to individual species. But how might their conclusions be important for broader planning, given that we have neither the resources nor the time to conserve species one-by-one⁹? Numerous species have overlapping distributions, such as the many taxa — maybe half of all the world's species — that co-occur in 25 'hotspots' that together make up only 1.4% of land area¹⁰. Data for these species are scarce, but patterns of habitat loss should match the geographies of their collapsing ranges¹¹. Based on Channell and Lomolino's results, we would expect disproportionate amounts of habitat to survive on the peripheries of the hotspots. This appears to be the case (Fig. 1)¹⁰. Should attempts to conserve these areas be concen-

trated on their edges? On average, the answer is 'no'³, because these geographically marginal areas will also tend to be ecologically marginal. Within the entire periphery of any one hotspot, though, there are likely to be areas that are particularly worth conserving, thanks to their isolation from the habitat loss driving extinction across the region.

Channell and Lomolino conclude that "the geography of recent extinctions" — by which they mean local extinctions — "is largely a geography of humanity". The geography of global extinctions, however, is also that of the hotspots¹². Within hotspots we cannot afford to ignore the lesson that peripheral sites, where biological diversity has been forced to live on the edge, may now hold some of the last opportunities for conservation. ■

Thomas Brooks is at the Zoological Museum, University of Copenhagen, the Department of Zoology, University of Cambridge, and the Center for Applied Biodiversity Science, Conservation International, 2501 M Street NW, Suite 200, Washington DC, 20037, USA.
e-mail: t.brooks@conservation.org

- Curnutt, J. L., Pimm, S. L. & Maurer, B. A. *Oikos* **76**, 131–144 (1996).
- Lomolino, M. V. & Channell, R. J. *Mamm.* **76**, 335–347 (1995).
- Lomolino, M. V. & Channell, R. *Conserv. Biol.* **12**, 481–484 (1998).
- Channell, R. & Lomolino, M. V. *Nature* **403**, 84–86 (2000).
- Gaston, K. J. *Ecography* **17**, 198–205 (1994).
- <http://www.creo.org>
- Olson, S. L. & James, H. F. in *Quaternary Extinctions: A Prehistoric Revolution* (eds Martin, P. S. & Klein, R. G.) 768–780 (Univ. Arizona Press, Tucson, 1984).
- Alroy, J. in *Extinctions in Near Time: Causes, Contexts, and Consequences* (ed. MacPhee, R. D. E.) 105–143 (Kluwer/Plenum, New York, 1999).
- Williams, P. H. in *Conservation in a Changing World* (eds Mace, G. M., Balmford, A. & Ginsberg, J. R.) 211–249 (Cambridge Univ. Press, 1998).
- Mittermeier, R. A., Myers, N., Robles Gil, P. & Goettsch Mittermeier, C. *Hotspots* (CEMEX, Mexico City, 1999).
- Brooks, T. M., Pimm, S. L. & Collar, N. J. *Conserv. Biol.* **11**, 382–394 (1997).
- Pimm, S. L. & Brooks, T. M. in *Nature and Human Society: The Quest for a Sustainable World* (ed. Raven, P.) 38–54 (Nat'l Acad. Press, Washington DC, 1999).

Apoptosis

Caspases find a new place to hide

Huseyin Mehmet

The philosophical muse proclaiming that “death is the essential condition of life”¹ has a firm footing in biology, for programmed cell death (apoptosis) is obligatory for normal development of multicellular organisms². But apoptosis is a double-edged sword, and deregulated cell death is implicated in a growing number of clinical disorders³. So, the apoptotic process needs to be tightly regulated. One way in which this is done is by physically segregating the different components of the apoptotic machinery — only when the death switch is tripped are the tools of execution brought together in the cytosol and the suicide programme activated.

The two main compartments known to be involved in such segregation are the plasma membrane, where both death and survival receptors reside, and the mitochondrion, which is home to several proteins that regulate apoptosis. Now, reporting on page 98 of this issue, Yuan and colleagues⁴ point to the endoplasmic reticulum as a third sub-cellular compartment implicated in apoptotic execution. Furthermore, they provide evidence linking activation of a hitherto obscure apoptotic enzyme, caspase-12, to Alzheimer's disease.

Famous mainly for the synthesis and processing of secreted proteins and the storage of intracellular calcium, the endoplasmic reticulum has already provided us with clues as to its alter ego. First, both pro- and anti-apoptotic members of the Bcl-2 family are located in intracellular membranes including the nuclear envelope, the outer mitochondrial membrane and the endoplasmic reticulum⁵. And second, cytosolic Ca²⁺ has been implicated as a pro-apoptotic second messenger⁶ involved in both triggering apoptosis and in regulating death-specific enzymes.

Among the most prominent of these death-specific enzymes is a family of cysteine-dependent aspartate-specific proteases known as the caspases⁷. These enzymes

can be broadly divided into two groups: initiator caspases (such as caspase-8 and caspase-9) whose main function is to activate downstream caspases, and executor caspases (such as caspases-3, -6 and -7), which are responsible for dismantling cellular proteins. The two main apoptotic pathways — the death receptor and mitochondrial pathways — are activated by caspase-8 and caspase-9, respectively, both of which are found in the cytoplasm (Fig. 1). Caspase-8 is recruited to a death-inducing signalling complex only when death receptors such as

Fas or the tumour-necrosis factor receptor are oligomerized after binding of specific ligands. In contrast, caspase-9 is activated when cytochrome c is released into the cytoplasm from the space between the inner and outer mitochondrial membranes.

Yuan and co-workers⁴ now show that another caspase, caspase-12, localizes not to the cytosol but to the endoplasmic reticulum. Caspase-12 is specifically involved in the apoptosis that results from stress in the endoplasmic reticulum. Treatment with compounds such as brefeldin A (which inhibits transport from the endoplasmic reticulum to the Golgi body) or tunicamycin (which inhibits N-glycosylation in the endoplasmic reticulum) triggers activation of caspase-12. But the strongest activation is seen in response to thapsigargin, which disrupts intracellular Ca²⁺ homeostasis, or to the Ca²⁺ ionophore A23187. Apoptosis triggered through pathways that do not involve the endoplasmic reticulum, such as serum deprivation or Fas activation, do not result in activation of caspase-12.

To study the localization of caspase-12, Yuan and colleagues used a specific caspase-12 antibody in immunoblotting experiments and detected bands in a fraction of brain extract that also contained an endoplasmic-reticulum-specific protein called TRAP α . Immunocytochemistry then revealed the perinuclear distribution of cas-

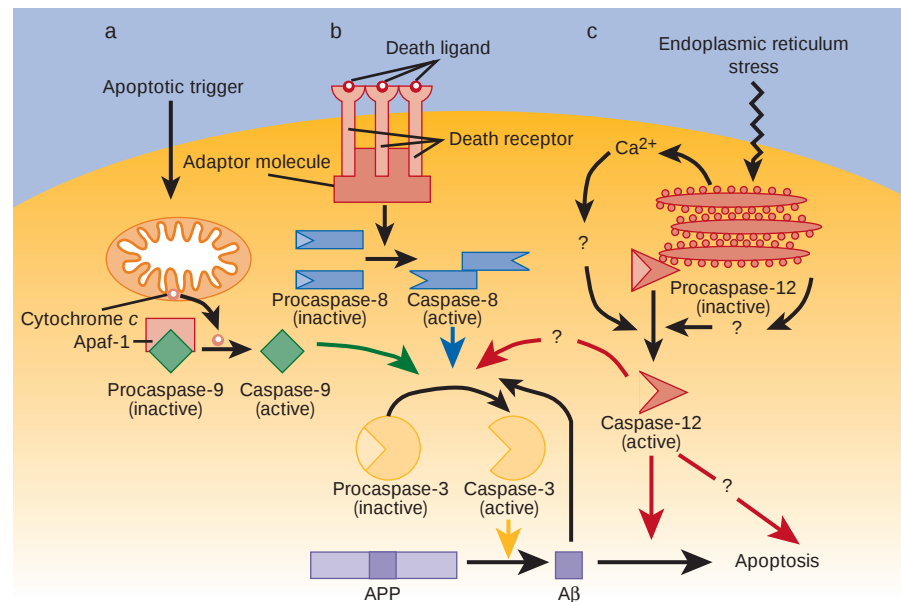


Figure 1 Three distinct apoptotic signalling pathways. a, When the mitochondrion receives appropriate apoptotic cues, or is irreversibly damaged, pro-apoptotic molecules such as cytochrome c are released into the cytosol. Together with ATP, cytochrome c forms a complex with Apaf-1 and procaspase-9, which is released in an active form. b, Oligomerization of death receptors (by specific death ligands) recruits adaptor molecules involved in activation of caspase-8. The active caspase is formed from two procaspase-8 molecules. c, After stress to the endoplasmic reticulum, including the release of Ca²⁺ from intracellular stores, caspase-12 is activated. Activated initiator caspases, such as caspase-8 and caspase-9, activate executioner caspases, including caspase-3. Active caspase-3 cleaves the β -amyloid precursor protein (APP), resulting in increased production of amyloid β -peptide (A β) which can feed back into caspase-3 activation and execute apoptosis in a caspase-12-dependent manner. The question marks denote possible, but unconfirmed, pathways.

pase-12, suggesting that it is situated in either the mitochondria or the endoplasmic reticulum. The co-localization of caspase-12 with several proteins found in the endoplasmic reticulum — grp78, presenilin-2, the β -amyloid precursor protein (APP) and green fluorescent protein targeted to the endoplasmic reticulum — confirmed that caspase-12 resides in the endoplasmic reticulum.

It seems, then, that the main function of caspase-12 is to facilitate apoptosis in cells irreversibly damaged by stress signals from the endoplasmic reticulum. Further evidence of this comes from Yuan and colleagues' studies of transgenic mice that lack intact caspase-12 protein. Thymocytes from these mice die in response to the mitochondria-dependent death signal dexamethasone, and hepatocytes die after injection of Fas-agonist antibodies, indicating that caspase-12 is not required for apoptosis initiated in the mitochondria or plasma membrane. In contrast, apoptosis of the renal tubular epithelium, which is mediated by endoplasmic reticulum stress, depends entirely on caspase-12 activity.

But this is not the end of the story. The endoplasmic reticulum, along with the associated presenilins and APP and the accumulation of Ca^{2+} , are all implicated in Alzheimer's disease⁸. Apoptosis has also been linked to Alzheimer's disease, with APP identified as a specific substrate for caspase-3 (ref. 9), and Yuan and colleagues' work now supports this idea. They found that primary cortical neurons isolated from their caspase-12-deficient mice are resistant to apoptosis induced by the cleavage product of APP, the amyloid- β peptide ($\text{A}\beta$). To exclude the possibility that the mutant mice lack a sub-population of $\text{A}\beta$ -sensitive neurons, the authors used an antisense approach to switch off caspase-12 in normal cells, and found that these cells were also protected against $\text{A}\beta$ toxicity.

Putting these data together we can envisage a model in which APP is cleaved by caspase-3, releasing $\text{A}\beta$ which, in turn, activates downstream caspases including caspase-3 (ref. 9) and, possibly, caspase-12 to amplify the effects. Although this needs further investigation, it does explain why, despite the high speed of apoptosis, Alzheimer's disease develops slowly over a period of years. Perhaps the presenilins and APP are innocent bystanders in neuronal cells that are triggered to undergo apoptosis after irreparable stress damage. The gradual accumulation of $\text{A}\beta$ could eventually reach critical levels, triggering neuronal death through a self-perpetuating circle of APP cleavage, $\text{A}\beta$ production and caspase-12-dependent apoptosis.

Like many important insights, Yuan and colleagues' experiments raise more questions than they answer (Fig. 1). For example, what are the physiological regulators of caspase-12? Although increases in the cytosolic

concentration of Ca^{2+} are enough to activate caspase-12, the protein intermediates in this process are not known. What are the downstream targets of caspase-12? Because caspase-3 is activated by $\text{A}\beta$, caspase-12 may be the missing link between APP cleavage, activation of caspase-3 and neuronal apoptosis. Where do Bcl-2 family proteins fit in? Their function is closely linked with caspase activity, and their localization in the endoplasmic reticulum suggests that they are involved in regulating caspase-12-mediated death.

If the identification of caspase-12 in the endoplasmic-reticulum stress pathway follows recent trends, we can expect there to be redundancy in the system. Indeed, the observations⁴ that such apoptosis is not completely inhibited in embryonic fibroblasts from the caspase-12-deficient mice suggests that there are caspase-12-independent pathways in the endoplasmic reticulum. But what sets caspase-12 apart from other apoptotic proteases is that it really does seem to be restricted to stress responses in the endoplasmic reticulum.

This observation has clear clinical ramifications. Caspases are central to both normal programmed cell death and injury-dependent apoptosis, so any therapy that manipulates caspase activity must take into account the possible effects on tissue homeostasis. In this regard, though, caspase-12 seems to

have a strong advantage as a target over other caspases. In contrast to other caspase knock-outs¹⁰, caspase-12-deficient mice have no noticeable developmental or behavioural defects, and have a normal incidence of tumours. So, caspase-12 is probably not essential in normal developmental death or tumorigenesis. And if activation of caspase-12 does turn out to be confined to only a narrow band of cellular stress signals, it will be a promising potential target for treating neurodegenerative diseases and cancer with few side effects. ■

Huseyin Mehmet is in the Weston Laboratory, Division of Paediatrics, Obstetrics and Gynaecology, Imperial College of Science, Technology and Medicine, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK.
e-mail: h.mehmet@ic.ac.uk

1. Gilman, C. P. *The Living of Charlotte Perkins Gilman: An Autobiography* (Appleton-Century, New York, 1935).
2. Jacobson, M. D., Weil, M. & Raff, M. C. *Cell* **88**, 347–354 (1997).
3. Thompson, C. B. *Science* **267**, 1456–1462 (1995).
4. Nakagawa, T. et al. *Nature* **403**, 98–103 (2000).
5. Krajewski, S. et al. *Cancer Res.* **53**, 4701–4714 (1993).
6. Nicotera, P. & Orrenius, S. *Cell Calcium* **23**, 173–180 (1998).
7. Wolf, B. B. & Green, D. R. *J. Biol. Chem.* **274**, 20049–20052 (1999).
8. Mattson, M. P., Guo, Q., Furukawa, K. & Pedersen, W. A. *J. Neurochem.* **70**, 1–14 (1998).
9. Gervais, F. G. et al. *Cell* **97**, 395–406 (1999).
10. Zheng, T. S., Hunot, S., Kuida, K. & Flavell, R. A. *Cell Death Differ.* **6**, 1043–1053 (1999).

Astronomy

A bright future for dark matter

Brad Hansen

Over the past 30 years, astronomers have accumulated an impressive body of evidence suggesting that there is more to our Galaxy than meets the eye. The motions of the stars we see betray the existence of a much larger gravitational pull than we can infer from the mutual interactions of the stars themselves. In fact as much as 90% of our Galaxy may be invisible to observers. This unseen mass has been dubbed 'dark matter' and is distributed roughly spherically throughout the Galaxy — mainly within the faint Galactic halo that surrounds the luminous Galactic disk. But the nature of this dark matter remains a complete mystery. On page 57 of this issue, Hodgkin *et al.*¹ present a small but vital piece of evidence supporting the notion that old dead stars may be an important constituent of dark matter.

Dark-matter theories fall into two broad classes. The first postulates that dark matter is not truly dark and resides in some kind of faint stellar object. Favoured candidates include brown dwarfs (an object about one-hundredth the mass of our Sun, too small to become a real star) and white dwarfs (a dead

star, the remnant of a star more massive than our own, which lived fast and died young).

The second class of theory is more speculative but has profound consequences if correct. In this class, the dark-matter candidates are various kinds of small elementary particles left over from the Big Bang. Such material was produced when the Universe was much hotter and denser, and would form the gravitational background of all subsequent evolution of stars and galaxies. These particles are much smaller than stars, but also more numerous, so that occasionally they will pass through the Earth and eventually, we hope, pass through subtle experiments designed to detect them. But until these experiments bear fruit, most progress on the dark-matter front will continue to focus on faint stellar objects.

In the past few years, gravitational microlensing experiments have opened a new avenue of investigation into this problem. These experiments monitored around ten million stars in the Magellanic Clouds (our nearest galactic neighbours), in the hope of detecting the occasional brightening

when something moves across our line of sight to one of the stars. This is because an object of stellar mass can act as a gravitational lens, focusing light and causing a transient increase in the observed flux. By measuring the duration of such a brightening, astronomers get a crude measure of the mass of the intervening lens.

Microensing experiments have provided two vital clues. The first is simply that they saw more events than expected from known stellar populations, supporting the idea that some of the dark matter in the Galactic halo consists of faint stellar objects. The second is that the durations of the events favour objects more massive than brown dwarfs², making white dwarfs the preferred candidates. But such inferences depend on which model you use, and other possibilities are allowed. In particular, microlensing events due to binary stars (which give extra velocity information) suggest that the Magellanic Clouds themselves may contribute a significant fraction of the lenses^{3,4}.

To settle this issue, we must try to find the proposed microlensing population by direct observation. For instance, white dwarfs are faint but not completely dark and we can hope to establish whether or not they exist in the Galactic halo if we look hard enough. In fact, we have already seen many white dwarfs in the Galactic disk, orbiting about the Galaxy in much the same way as our Sun. But we believe that the Galactic halo is older than the Galactic disk, and so any white dwarfs contributing to the dark matter should be older and fainter than the known population.

The problem then becomes "What do the oldest white dwarfs look like?". Previous searches for stars this old were unsuccessful because of the assumption that they were looking for cool, and therefore red, objects. Current theory suggests that very old white dwarfs should appear surprisingly blue, because of strong absorption of radiation at infrared wavelengths by molecular hydrogen in their atmospheres^{5,6}. Tantalizing observa-

tions by the Hubble Space Telescope⁷ suggest that some objects in the Hubble Deep Field survey have proper motions and colours that are consistent with what we now expect for old white dwarfs. But these objects are very faint and the colours provide only crude measures of their physical properties. Are they truly white dwarfs? What we need is a well-characterized object to guide our searches.

That is exactly what Hodgkin *et al.*¹ provide. They describe the properties of an extremely cool (and therefore old) white dwarf, WD0346 + 246, which has a known distance and velocity. The remarkable thing about this object is the radiation it emits at long wavelengths. The flux emerging in the infrared is dramatically suppressed, supporting our ideas about how white dwarfs change character as they age. Furthermore, the velocity of this object indicates that it belongs to the spherical Galactic halo, as opposed to most known white dwarfs, which orbit the Galactic centre in the planar Galactic disk. Hodgkin *et al.* estimate that the probability that this white dwarf belongs to the known stellar population is small — that is, it is probably the first example of a white dwarf that might contribute to the spherically distributed dark matter. Of course, this is only one object, so some caution is called for. Nevertheless, at the very least it shows that we are on the right track and may continue our searches with greater confidence now that we know what we're looking for. A more optimistic interpretation would be that they have found the tip of the iceberg. ■

Brad Hansen is in the Department of Astrophysical Sciences, Princeton University, Princeton, New Jersey 08544-1001, USA.

e-mail: hansen@astro.princeton.edu

1. Hodgkin, S. *et al. Nature* **403**, 57–59 (2000).
2. Alcock, C. *et al. Astrophys. J.* **486**, 697–726 (1997).
3. Bennett, D. P. *et al. Nucl. Phys. B (Proc. Suppl.)* **51**, 131 (1996).
4. Alcock, C. *et al. Astrophys. J.* **518**, 44–49 (1999).
5. Hansen, B. M. S. *Nature* **394**, 860–862 (1998).
6. Saumon, D. & Jacobsen, S. B. *Astrophys. J.* **511**, L107–L110 (1999).
7. Ibata, R. A., Richer, H. B., Gilliland, R. L. & Scott, D. *Astrophys. J.* **524**, L95–L98 (1999).

Earth science

Taking the temperature of slabs

Stephen H. Kirby

Subduction zones, where oceanic plates dive deep into Earth's mantle, produce graceful arc-shaped chains of volcanoes landward of oceanic trenches and parallel belts of earthquakes marking plate descent. Sinking plates (slabs) preserve the cold structure of Earth's near-surface thermal boundary layer to great depths. They also carry minerals of the shallow oceanic crust and mantle to depths where they

become unstable, transform into denser forms, and release fluids. Such metamorphism of slab rocks is central to the processes that produce volcanic arc magmas and to those that permit the occurrence of earthquakes deeper than a few tens of kilometres where pressures become too large.

Papers in *Geophysical Research Letters*¹ and *Science*² by Peacock and Hyndman, and Peacock and Wang, provide insights into the

effects of variable thermal structure on the stability of slab minerals. By integrating thermal modelling of subducting plates with information on mineral stabilities, the authors provide support for recent models of both subduction-zone volcanism and earthquakes^{3,4}. These new models give quantitative insight into the critical depths over which slabs release water stored in hydrated minerals, and the effects of this water on seismic and volcanic processes. The authors emphasize differences in thermal structure and in the depths of water release between normal 'cold' subduction and the less-common 'warm' subduction. Old plates have a long time to cool after their birth at mid-ocean ridges and, if they descend as slabs at fast rates, they stay cold to great depth. Young plates lose less of their original heat and tend to descend at slower rates, and so are much warmer.

Peacock and Hyndman¹ explore the effects of water on the depth limit of rapid seismic slip along the boundary between the sinking slab and the upper plate. They consider the effects of water released from slabs on the mineralogy of the nearby upper plate (or forearc). For normal cold subduction (Fig. 1a, overleaf), appreciable interplate earthquake slip usually occurs to depths of about 40 km (refs 5, 6). This depth limit roughly corresponds to the depth of the crust–mantle boundary in the forearcs of many subduction zones⁷. The shallow part of the subducting crust loses water by compaction and by dehydration of low-grade minerals during descent. This water could hydrate the cold forearc mantle above the slab, producing the rock serpentinite. Serpentinite minerals tend to exhibit stable sliding behaviour in laboratory friction experiments, and so a hydrated forearc could make that part of the plate boundary aseismic³.

From their calculated pressures and temperatures, Peacock and Hyndman show that the serpentinite minerals antigorite and brucite are indeed thermodynamically stable in the forearc mantle. Both minerals show stable-sliding behaviour in the laboratory^{8–10}. Peacock and Hyndman also argue that water liberated from the descending slab should transport silica dissolved in the aqueous fluids to the mantle above, and that a layer of talc should form near the sliding surface by reactions of silica with antigorite and brucite. Talc shows stable sliding frictional behaviour, at least at room temperature. Thus all three minerals that are expected to occur in this setting should exhibit stable sliding behaviour. Slip between slabs and hydrated forearc mantle should therefore tend to be seismically silent, as is often observed.

Earlier thermal models for warm slabs by Hyndman, Wang and co-authors^{3,11} show that isotherms at the sliding plate boundary are markedly shallower than those in cold slabs (Fig. 1b). The depth limit of sudden

seismic slip occurs at temperatures of 350–450 °C at 15–25 km, above the crust–mantle boundary in many forearcs. Hyndman *et al.*³ associate this shallower transition to aseismic slip with the onset of stable frictional sliding of quartzo-feldspathic rocks that are said to make up the shallow crust of the forearc.

In the second paper, Peacock and Wang² extend investigation of these differences in behaviour between cold and warm slabs to deeper phenomena. Warm subduction (Fig. 1b) usually produces mostly shallow intraslab earthquakes (up to 70-km depth), feeble or absent volcanic activity, and lava chemistries that can differ from those of typical arc volcanoes^{2,4}. Their models² show that the youngest (15–27-million-year-old) crust

of the Philippine Sea plate that subducts beneath southwest Japan is as much as 300–500 °C warmer than the older crust of the Pacific plate subducting beneath northeast Japan. Such warm mafic crust (basalts and gabbros and their hydrous metamorphic equivalents) should dehydrate and transform into the rock eclogite at depths of only 40–90 km. Hydrous mafic rocks lining faults in crust created beneath the ocean basins can be reactivated during slab dewatering by producing high internal water pressure, thereby reducing the shear stresses required to overcome rock friction⁴. Earthquakes caused by this mechanism should therefore be limited to the shallow depths over which slab dewatering is expected, as observed.

Moreover, these warm-slab models pre-

dict that the principal dewatering of the subducting mafic crust should occur before it descends below the asthenosphere (the hot, weak and chemically fertile source of magmas in the mantle beneath the lithosphere). Melts produced by fluxing the asthenosphere with water should therefore be less abundant than in cold-slab settings. Melt production at depth may be limited to high-temperature melting of the subducted eclogitic crust. Thus the production rates of arc volcanoes should be low and their lavas can have the distinctive geochemistry of melted eclogite — also as observed.

The pressure–temperature paths for mafic crust in colder slabs (such as that beneath northeast Japan) pass through the lawsonite–blueschist metamorphic facies on their way down to the eclogite field (Fig. 1a). Water stored in these blueschist minerals should persist in bound form to depths of more than 100 km. Although the precise positions of the boundaries defining the deep limits of this blueschist mineral assemblage are not known experimentally because of sluggish reaction kinetics, thermodynamic models suggest that these minerals should be stable to depths of over 100 km in cold slabs. A large fluid loss from the subducting oceanic crust could therefore occur at depths of 100–200 km, where normal arc magmas are thought to originate and where most earthquakes of intermediate depth occur.

These models^{1–4} are no doubt oversimplifications. Nonetheless they provide self-consistent interpretations of the depth distributions of interplate and intraslab earthquakes, the distribution of arc volcanic activity, and the gross chemistries of arc magmas. Improvements are needed to establish the precise equilibrium thermodynamic stability fields of blueschist rocks, as are reaction-rate studies of eclogite-forming reactions to help show whether such reactions keep pace with slab descent or not. Most thermal models, including those discussed here, also assume that thermal conductivity and diffusivity are constant for both the subducting crust and mantle. These heat-transfer properties of slab rocks and minerals need to be measured under likely slab conditions to see if this assumption leads to serious errors in calculated slab temperatures.

There may also be a practical side to this line of research. For example, many shallow intraslab earthquakes of magnitude 7–8 have occurred in this century beneath populated regions of Mexico, the Puget lowlands of the northwest United States, and southwest Japan, all sites of warm-slab subduction. Hazard assessment is especially difficult in these circumstances, because the causative faults are not exposed on the Earth's surface. Their long-term fault-slip histories are therefore unknown and their deformations may be too small to measure at the Earth's surface. These models^{2,11} for

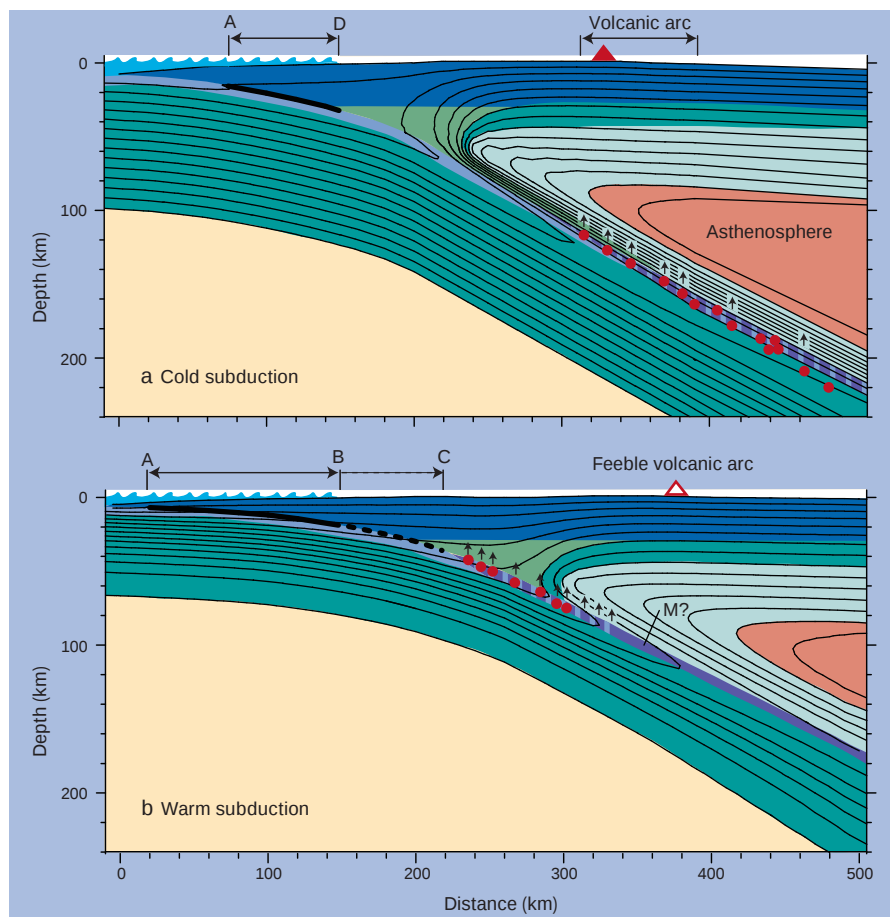


Figure 1 Models for slab subduction, adapted from refs 1–4. Contour interval is 100 °C. a, Cold subduction. Interplate seismic slip is shifted to greater depths (to the interval A–D) than in warm subduction (where subducting crust is as much as 300–500 °C hotter). Slip, in turn, becomes aseismic near the point of slab contact (D) with serpentinites (light green) of the forearc mantle which are made up of minerals that exhibit stable frictional sliding in the lab. Colder temperatures also mean that dehydration (arrows) of the subducting lower crust happens at depths exceeding 100 km, leading to both deeper intraslab earthquakes (red dots) and normal arc volcanism produced by partial melting of the asthenosphere. **b, Warm subduction.** A hotter thermal structure restricts the region of seismic slip on the plate contact to shallower depths (A–B or A–C), but over a larger area, producing potentially large offshore interplate earthquakes. Water released from the lower crust at depths of 40–90 km reflects shallow metamorphism of normal mafic crustal rocks (light blue) to eclogite (purple) and facilitates shallow intraslab earthquakes (red dots). Shallow dewatering should also reduce partial melting of the asthenosphere and thus reduce normal arc volcanic activity. Higher temperatures may also cause limited slab crustal melting (M?) and produce anomalous chemistries in the lavas of the associated volcanoes.

warm subduction provide a thermophysical framework for future studies of the seismic component of deformation rates inside warm slabs.

Stephen H. Kirby is at the US Geological Survey, MS 977, 345 Middlefield Road, Menlo Park, California 94025, USA.

e-mail: skirby@isdmnl.wr.usgs.gov

1. Peacock, S. & Hyndman, R. *Geophys. Res. Lett.* **26**, 2517–2520 (1999).
2. Peacock, S. & Wang, K. *Science* **286**, 937–939 (1999).

3. Hyndman, R., Yamano, M. & Oleskevich, D. *The Island Arc* **6**, 244–260 (1997).
4. Kirby, S., Engdahl, E. & Denlinger, R. *Am. Geophys. Un. Geophys. Monogr.* **96**, 195–214 (1996).
5. Zhang, Z. & Schwartz, S. J. *Geophys. Res.* **97**, 537–544 (1992).
6. Tichelaar, B. & Ruff, L. J. *Geophys. Res.* **98**, 2017–2037 (1993).
7. Ruff, L. & Tichelaar, B. *Am. Geophys. Un. Geophys. Monogr.* **96**, 105–111 (1996).
8. Reinen, L. *et al.* *Pure Appl. Geophys.* **143**, 317–358 (1994).
9. Moore, D. E. *et al.* *J. Geophys. Res.* **102**, 14787–14801 (1997).
10. Moore, D. E. *et al.* *Trans. AGU Fall 1997 Mtg Eos* **78**, F732–F733 (1997).
11. Wang, K., Hyndman, R. & Yamano, M. *Tectonophysics* **248**, 53–69 (1995).

Cell biology

Telomeric tethers

Wai-Hong Tham and Virginia A. Zakian

Eukaryotic chromosomes are confined to the nucleus, which is separated from the rest of the cell by two concentric membranes known as the nuclear envelope. The nuclear envelope is punctuated by holes known as nuclear pore complexes, which provide the main pathway for transport of cellular material across the nuclear–cytoplasmic boundary. The nuclear pore complex is a complicated structure containing up to 100 or more proteins called nucleoporins. The data presented by Galy *et al.*¹ on page 108 of this issue suggest that the three-dimensional organization of chromosomes is provided, in part, by interactions between nucleoporins and specific chromosomal regions.

Telomeres — the protein–DNA structures at the ends of chromosomes — are essential for the stable maintenance of eukaryotic chromosomes. In some organisms (including yeast), telomeres also repress the transcription of nearby genes, a phenomenon known as telomeric silencing or telomere-position effect². In the yeast *Saccharomyces cerevisiae*, telomeric DNA consists of 300 ± 75 base pairs of C₁₋₃A/TG₁₋₃ repeats. Adjacent to these simple repeats is Y', a longer DNA element that is found at about two-thirds of yeast chromosome ends.

At least seven structural proteins bind yeast telomeres *in vivo*^{3,4}, including Rap1p (the main structural protein at telomeres), three Sir proteins and the Yku complex, which is a heterodimer of the Yku70p and Yku80p proteins. All of these proteins are concentrated in foci close to the nuclear periphery^{5–7}, and each is required for telomeric silencing. Using *in situ* hybridization, about 70% of the Y' DNA has been shown to co-localize to these foci of telomeric proteins⁵. Each diploid yeast cell has 64 telomeres but only six to seven foci of Y' DNA, indicating that many yeast telomeres are clustered near the nuclear periphery. Although mutations that eliminate the foci of silencing proteins do not necessarily change the localization of Y' DNA^{5,8}, telomere localization and

telomeric silencing are disrupted in strains that lack Yku70p (ref. 7). The functional significance of locating telomeres at the nuclear periphery is not known; however, sequestering silencing proteins in a limited number of foci may protect the rest of the genome from transcriptional silencing.

Galy *et al.*¹ now show an intriguing link between Ku proteins, nuclear pore complexes and telomere localization. They studied two proteins associated with the nuclear envelope, Mlp1p and Mlp2p. Mlp1p was initially identified in a screen for nuclear-envelope-associated proteins that are not part of the nuclear pore complex. Mlp1p and the highly related Mlp2p are concentrated near the nuclear envelope on the side facing the inside of the nucleus⁹.

Galy and colleagues have found that if they delete the *MLP2* gene, localization of Yku70p is perturbed. However, this strain shows normal telomere-position effect, suggesting that altered localization of Yku70p is not enough to cause a reduction in telomeric silencing. The authors also showed that Mlp2p co-immunoprecipitates with Yku70p, indicating that these two proteins

interact *in vivo*. Deletion of both *MLP1* and *MLP2* alters the localization of Y' DNA. The double mutant has more Y' foci than normal, suggesting a loss of telomere clustering. In addition, compared with wild-type cells, these foci were less likely to be located at the nuclear periphery. Inexplicably for proteins involved in telomere localization, the Mlp proteins (as well as the Yku complex¹⁰) are involved in double-strand-break repair, as loss of either confers sensitivity to the radio-mimetic drug bleomycin.

To date, only the Yku complex has been found to affect the nuclear localization of telomeres^{6,7}. But, together with previous observations, Galy and colleagues' results lead to a model in which the Mlp proteins mediate the peripheral localization of telomeres by binding to telomere-associated Yku70p (Fig. 1). To pursue this result, the authors isolated other, unlinked mutations that caused a yeast strain lacking the *MLP1* gene to die. Using this approach they identified a gene called *NUP145*, which encodes an essential nucleoporin involved in transport of polyadenylated RNAs¹¹. They found that a strain (*nup145ΔC*) producing a variant of Nup145p that lacks the carboxy terminus of the protein results in a striking mislocalization of the Mlp proteins. The authors also observed mislocalization of the telomere-binding protein Rap1p, and an increased number of Rap1p foci in the *nup145ΔC* strain.

Based on these observations, Galy and colleagues suggest that the Mlp proteins are responsible for the peripheral localization of telomeres, and that Mlps are docked to the nuclear periphery by an interaction with Nup145p (Fig. 1). However, these data present a paradox. Other groups have shown that the absence of a different nucleoporin, Nup133p, causes nuclear pore complexes to cluster to a sub-region of the nuclear envelope, yet in this case the localization of Rap1p is normal⁵. This result is not expected if nuclear pore complexes mediate telomere

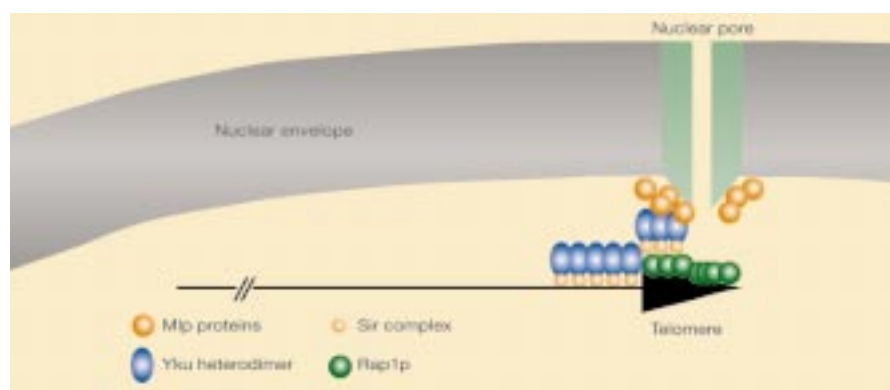


Figure 1 Model for telomere localization. Galy *et al.*¹ propose that the Mlp2 protein helps localize telomeres to the nuclear periphery by binding to Yku70p, a telomere-binding protein that exists as a heterodimer with Yku80p. This complex is thought to dock to the nuclear envelope by an interaction between Mlps and the nuclear pore complexes. The Sir complex and Rap1p are additional telomere-binding proteins.

localization. This contradiction raises questions about the plasticity of nuclear pore complexes that can be addressed by localizing Nup145p under conditions where nuclear pore complexes are clustered but telomeric silencing is normal.

There are additional caveats. Several conditions have been described^{5,8} in which the foci of telomeric proteins are dispersed, telomeric silencing is lost, yet Y' DNA is still localized to the nuclear periphery. Because Galy and colleagues monitored the localization of Rap1p rather than Y' in the *nup145ΔC* strain, it will be important to establish that telomeric DNA — not just telomeric proteins — are perturbed in the absence of wild-type Nup145p. In addition, to assess the relationship between telomeric silencing and telomere localization, the position of individual telomeres whose silencing status is known must be determined. Finally, conclusions as to the loss of telomeric silencing of the *Δmlp1*, *Δmlp2* and *nup145ΔC* strains should be interpreted with caution as these strains grow slowly in complete media^{9,11}.

Nonetheless, Galy and colleagues provide the first hint of a possible link between telomere localization and nuclear pore complexes. Future experiments with MLps should bring further insights into the connection between telomeric silencing, telomere localization and DNA repair. ■

Wai-Hong Tam and Virginia A. Zakian are in the Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544, USA.
e-mails: waitham@phoenix.princeton.edu
vzakian@molbio.princeton.edu

1. Galy, V. *et al.* *Nature* **403**, 108–112 (2000).
2. Gottschling, D. E. *et al.* *Cell* **63**, 751–762 (1990).
3. Bourns, B. D., Alexander, M. K., Smith, A. M. & Zakian, V. A. *Mol. Cell. Biol.* **18**, 5600–5608 (1998).
4. Gravel, S. *et al.* *Science* **280**, 741–744 (1998).
5. Gotta, M. *et al.* *J. Cell Biol.* **134**, 1349–1363 (1996).
6. Gotta, M. *et al.* *EMBO J.* **16**, 3243–3255 (1997).
7. Laroche, T. *et al.* *Curr. Biol.* **8**, 653–656 (1998).
8. Martin, S. G. *et al.* *Cell* **97**, 621–633 (1999).
9. Strambio-de-Castillia, C. *et al.* *J. Cell Biol.* **144**, 839–855 (1999).
10. Haber, J. E. *Cell* **97**, 829–832 (1999).
11. Teixeira, M. T. *et al.* *EMBO J.* **16**, 5086–5097 (1997).

Methane combustion

Durable catalysts for cleaner air

Jon G. McCarty

There is great interest in using methane as an alternative to coal and oil, because its high ratio of hydrogen to carbon means it can generate power with fewer greenhouse-gas emissions. But methane combustion requires high temperatures (greater than 1,300 °C) to maintain a stable flame. At these temperatures, noxious emissions of nitrogen oxides that contribute to acid rain and smog become a problem. Catalytic oxides can be used to stabilize 'lean' flames (those with low fuel to air ratio) at much lower temperatures, thereby producing little or no nitrogen oxide pollutants. On page 65 of this issue¹, Zarur and Ying describe the synthesis of promising new catalytic materials that allow methane combustion at 400 °C.

The catalytic reaction of methane and oxygen was examined as long ago as 1817 (before the term catalysis was first coined by Berzelius²) when Humphry Davy and young Michael Faraday investigated the ability of platinum to spontaneously release heat when exposed to mixtures of coal-mine gas and air³. Apart from early work on safety devices for coal mines and the burning of volatile organic vapours, the catalytic oxidation of fuel was not investigated for commercial applications until the mid-1970s when NASA and the Environmental Protection Agency in the United States sought to develop a small ceramic gas turbine as a low-emissions alternative to internal combustion engines for light vehicles⁴. By the early 1980s, catalytic exhaust converters were available to meet new vehicle-emission standards, but the durability and performance of the ceramic gas turbine remained unacceptable, despite progress in the development of catalytic materials.

Attention has since shifted, especially in Japan, to low-emission catalytic and hybrid combustion for stationary electrical-power generation with large-scale (megawatts), natural-gas turbines. Current catalytic combustion devices for small- to medium-scale (0.5- to 15-MW) gas turbine power-generation systems use a very lean mixture of natural gas with compressed air. Our company, Catalytica Combustion Systems, Inc., is developing catalytic combustion technology that is both durable and cost-effective relative to exhaust clean-up. To demonstrate commercial viability, Catalytica has applied this technology to a 1.5-MW gas turbine, which supplies power to the city of Santa Clara, California. To date, the catalyst has run for more than 4,000 hours with very low emissions of nitrogen oxides, carbon monoxide and unburned hydrocarbons (Fig. 1, overleaf). Still, it is fair to say that commercial acceptance of this technology has been slowed by the limited stability of solid-state catalysts.

Catalytic combustion materials have to be hard wearing because they are required to operate for long periods without loss of activity at high temperatures. Noble metal oxides, such as platinum or palladium oxide, are excellent catalysts below 850 °C, but powders of these catalysts suffer from thermal coarsening at higher temperatures. Thermal coarsening is the growth of larger grains of catalyst material through the consumption of smaller grains, which reduces the catalyst surface area and thus its activity. To improve palladium oxide activity at higher operating temperatures the catalyst must be dispersed within a more stable (and typically less active) porous oxide layer (Fig. 1b). This

Reproductive biology

Taking the squeeze off sperm

Non-hormonal contraception is available for males — but, for the moment, only for mice and by targeted gene deletion. The target is the gene encoding a receptor known as P2X₁, which is prominent in smooth-muscle cells of the vas deferens, the ducts that carry sperm from the testicles *en route* to ejaculation. The image here shows a cross-section of vas deferens brimming with P2X₁.

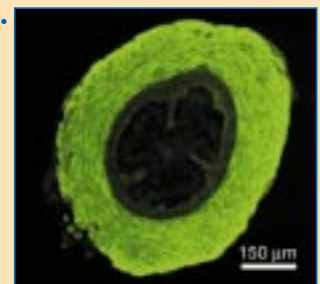
Elsewhere in this issue (*Nature* **403**, 86–89; 2000), K. Mulryan and

colleagues describe the results of knocking out the P2X₁ receptor gene in mice. Females mated with male mice lacking the receptor had a pregnancy rate of only 13.7% against 100% for females mated with normal males.

The cause of the infertility is a much reduced sperm count. Nerve stimulation of vas deferens muscle is largely through P2X₁, and usually results in vas deferens contraction and sperm ejection. In the knockout mice, it seems that the absence of

P2X₁ means that stimulation fails to reach the threshold for normal contraction.

For human contraception, ways of disrupting P2X₁ other than gene deletion would have to be found. A further possible difficulty is that these receptors occur widely elsewhere in the body, and interfering with their function might have side effects. The knockout mice of Mulryan *et al.* showed no obvious abnormalities in behaviour and physiology, however, except for



marginally raised blood pressure.

There is also the other side of the fertility coin. As the authors point out, stimulation of P2X₁ activity might prove fruitful in treating male infertility.

Tim Lincoln

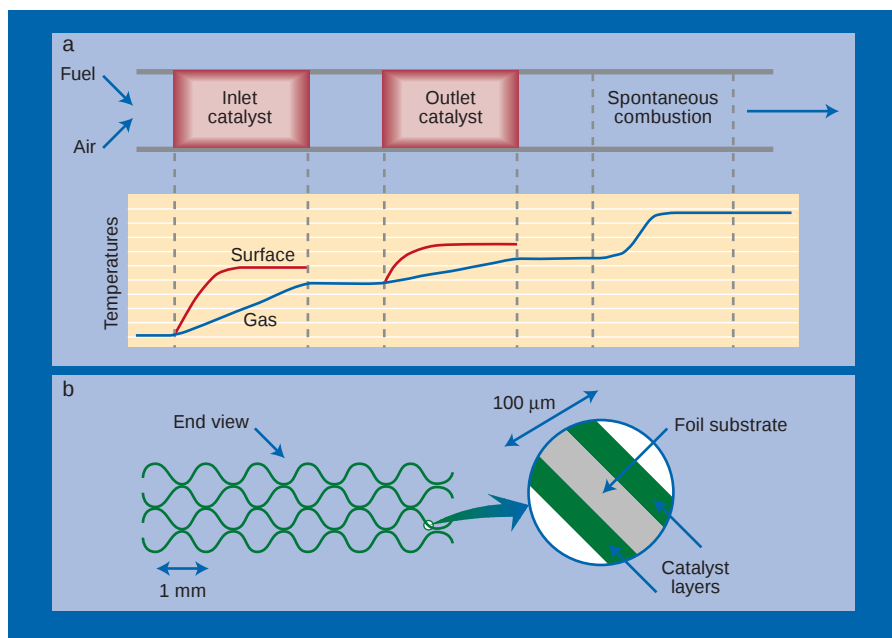


Figure 1 Catalytic combustion technology for gas-turbine applications⁶. a, Fuel and compressed air pass through two catalytic stages, in which catalyst materials limit the temperature rise through each stage. The gas exiting the second stage contains unburned fuel and is sufficiently hot to spontaneously combust in the post-catalyst zone, just before entering the gas turbine. b, An end view of the catalyst honeycomb structure. Thin strips of metal foil are coated with a layer of catalyst and configured to create an array of narrow (1-mm) channels. Fuel in the gas mixture diffuses through the layers where the catalytic oxidation occurs, thereby releasing heat. The heat of combustion is then returned to the flowing gas by convection. Finding catalysts that work at the high temperatures required for methane combustion is not easy. Zarur and Ying¹ have developed a promising catalyst that is more stable at temperatures up to 1,300 °C and could be used in a variety of combustion applications.

prevents active particles meeting and fusing together, thereby preserving the high surface area of the catalyst. The activity of such catalysts can be viewed as the product of their intrinsic ability to oxidize methane (rate per unit surface area of the active phase) and their dispersion (active surface area per unit volume). It is therefore desirable to find materials that combine high activity with greater resistance to thermal coarsening and so maintain high combustion rates.

Oxide catalysts based on non-noble metals, such as perovskites and hexaaluminates, are exceptionally resistant to thermal coarsening, but are difficult to prepare as fine powders. Zarur and Ying¹ have found a way round this problem by combining several state-of-the-art synthesis methods to produce porous powders with high surface area — that is, nanometre particle sizes — and resistance to thermal coarsening at high temperatures. These hexaaluminate powders show an interesting level of activity for methane combustion, but the most important use of these materials may be as a matrix to disperse and isolate small particles of other catalysts that are more prone to coarsening. Indeed, Zarur and Ying¹ disperse grains of a moderately active catalyst, ceria (CeO₂), within a microemulsion of their hexaaluminate oxide powders with interesting results. In this system, the highly dis-

persed and surprisingly efficient ceria catalyst helped sustain methane combustion at temperatures as low as 400 °C.

The dispersed ceria catalyst appears to remain stable and active within the limited conditions and lifetime of laboratory experiments, and is perhaps superior to the transition-metal hexaaluminates studied by Arai and colleagues some ten years ago⁵. The new catalyst is perhaps unique in that ceria, a low-volatility oxide previously thought to have low combustion activity, remains active and highly dispersed within the porous hexaaluminate support after extended exposure to combustion conditions at temperatures as high as 1,300 °C. It remains to be seen if this catalyst possesses the combination of thermal stability and combustion activity necessary to succeed as a commercial catalytic combustion technology. But clearly the ceria-hexaaluminate catalyst, or another of many possible variations prepared using this approach, is worth a closer look. ■

Jon G. McCarty is at Catalytica Combustion Systems, Inc., 430 Ferguson Drive Building 3, Mountain View, California 94043-5272, USA. e-mail: jmccarty@mv.catalytica-inc.com

1. Zarur, A. J. & Ying, J. Y. *Nature* **403**, 65–67 (2000).
2. Berzelius, J. J. *Ann. Chim.* **61**, 146 (1836).
3. Davy, H. *Phil. Trans. R. Soc.* **107**, 45 (1817).
4. Pfeifferle, W. *J. Energy* **2**, 142–146 (1978).
5. Arai, H. & Machida, M. *Catal. Today* **10**, 81–95 (1991).
6. Dalla Betta, R. A. *Catal. Today* **35**, 129–135 (1997).

Daedalus

Depth of insight

Computed axial X-ray tomography is an expensive and cumbersome way to obtain three-dimensional X-ray data. Daedalus wants to do it by stereo photography. In its optical form, this uses two lenses separated by the distance between the eyes; each takes a photograph. These are then presented to the viewer so that each eye can only see its 'own' photograph. The result is a persuasive and most informative image in depth. The method should work perfectly well with two X-ray sources, or just one stepped quickly sideways between successive exposures.

The snag is that an X-ray image is essentially a shadowgraph. A point source of X-rays throws a shadow of the subject, in all its variations of transparency or opacity, onto a sensitive screen. With two sources, the screen would just sum the two exposures, giving a confused muddle. An X-ray stereography screen will have to hold a separate image for each of the sources.

Daedalus recalls those fine filters made by exposing a sheet of base material to a beam of ionizing particles. Each particle leaves a track through the sheet, dense with dislocations, photoelectrons and so on. A selective solvent etches away these disturbed regions, leaving a tiny tunnel where each particle passed.

So DREADCO chemists are now seeking a polymer or light alloy which can be activated and etched by the passage of X-ray photons. It will be placed behind the subject; in front of him will be two X-ray sources, separated by the width between a human interpreter's eyes. Each X-ray source will riddle the plate with tracks aligned with it, and diverging from it. The interpreter will etch the plate, place it against a bright background, and look at it from such a position that his two eyes are where the X-ray sources were. Each eye will see only the light which converges on it through the tunnels in the plate made by the corresponding X-ray source, and aligned with it. He will see a true stereo X-ray image of the subject.

Cheap and simple, stereo X-radiography will transform routine medical imaging. Each exposure will give far more useful data; many tricky interpretational problems will vanish. It could even provide airports with still more intimate insights into their customers' luggage.

David Jones

The further Inventions of Daedalus (Oxford University Press), 148 past Daedalus columns expanded and illustrated, is now on sale. Special *Nature* offer: m.curtis@nature.com

Marine iguanas shrink to survive El Niño

Changes in bone metabolism enable these adult lizards to reversibly alter their length.

Change in body length is considered to be unidirectional in vertebrates¹, but we have repeatedly observed shrinkage in the snout-to-vent length of individual adult iguanid lizards. In two studies, one lasting 18 years and one 8 years, of two island populations of Galápagos marine iguanas (*Amblyrhynchus cristatus*), we found that individuals became shorter by as much as 20% (6.8 cm) within two years. This shrinking coincided with low availability of food, resulting from El Niño events. Body length increased again during subsequent La Niña conditions, when algal food was abundant². We found that lizards that shrank more survived longer than larger iguanas during harsh periods because their foraging efficiency increased and their energy expenditure decreased^{3,4}.

Marine iguanas (Fig. 1) are herbivorous reptiles that feed on submerged intertidal and subtidal algae along the rocky island shores of the Galápagos archipelago, Ecuador³. Environmental conditions can change dramatically within an individual's lifetime, as they can live for more than 28 years². El Niño events usually recur at intervals of 3–7 years, but were more prevalent between 1990 and 1999: the El Niño Southern Oscillation Index, a composite measure of the severity of El Niño, was positive throughout 1990–95 and 1997–98 (ref. 5). During El Niño years in the Galápagos, sea surface temperatures increase from an average of 18 °C to a maximum of 32 °C as cold ocean currents and cold nutrient-rich upwellings are disrupted⁶.

The increased rainfall that accompanies El Niño results in greater food availability for most terrestrial organisms in the Galápagos⁷, but marine life generally suffers⁸. Green and red algal species, which are the marine iguanas' preferred food, disappear and are replaced in intertidal areas by brown algae⁸ which iguanas find hard to digest. Up to 90% of marine iguana populations on islands can die of starvation as a result of these environmental changes⁹. The largest iguanas have the highest mortality because, as is often the case with grazers¹⁰, they feed less efficiently than smaller individuals.

During the most recent El Niño event (1997–98), larger individuals of the two island populations shrank more than smaller individuals (even if calculated as a percentage), and females shrank more than males of the same size (Fig. 2a). This sex-specific shrinkage is presumably related to females exporting energy into clutches of eggs during the year before an El Niño³. The scale of the shrinkage — up to 20% of body length —



Figure 1 The marine iguana *Amblyrhynchus cristatus*.

means that it cannot simply be explained by decreases in cartilage and connective tissue, which together make up only 10% of total body length¹¹. We believe that bone absorption accounts for much of the reduction.

Shrinking in marine iguanas may be an adaptive response to low food availability and energetic stress. Measurements of a cohort of adults more than 300 mm long during the strong 1992–93 El Niño event^{2,3,5} show that individuals that shrank more survived significantly longer (Fig. 2b). The mechanisms that determine whether and to what extent an individual shrinks during El Niño events remain unclear.

Reduction in body length has been observed previously, and growth rates set to zero by definition¹, but to our knowledge this is the first report of shrinkage in adult vertebrates. In humans, the long bones of astronauts become reduced both in density and length during prolonged space trips¹². Astronauts experience long periods of microgravity, resulting in limited exercise and high cortisol ('stress') levels^{13,14}, and confining healthy volunteers to months of bed rest has much the same effect¹². Similarly, marine iguanas get very little exercise from feeding during El Niños⁴ and their corticosterone levels are very high (L. M. Romero and M. W., manuscript in preparation). There may be a causal relationship between limited exercise, high corticosterone levels and decreases in body length¹⁴. Similar relationships have been suggested for the regulation of human osteoporosis¹⁵.

Adult marine iguanas can switch repeatedly between growth and shrinkage during their lifetime, depending on environmental conditions². In humans, bone regrowth after osteoporosis is largely impossible, causing major health problems (for example, in post-menopausal women¹⁵). It is not clear whether other vertebrates can shrink and regrow. We propose that growth rates should no longer be set to zero, as shrinkage can confer a biologically important advantage.

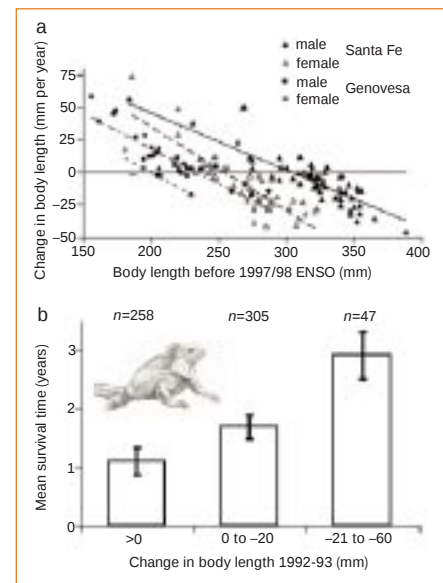


Figure 2 Change in body length as an adaptation to harsh conditions. **a**, Reduction in body length (snout-to-vent length, SVL) of Galápagos marine iguanas on the islands of Santa Fe (90° 02' W, 0° 50' S) and Genovesa (89° 59' W, 0° 19' N), Ecuador. Males (filled symbols) shrink less at a given body length than females (open symbols). Individuals are marked for life⁸ and measurements were taken without knowledge of shrinkage pattern or previous body length. Mean ($\pm$ s.d.) maximum intra- and inter-individual measurement errors were $1.47 \pm 0.6\%$ and $2.1 \pm 0.8\%$ body length ($n=9$ and 12), respectively. Body mass and condition do not influence SVL measurements. Lines show linear regressions for each island and sex. Santa Fe males: shrinkage, $134(\pm 12) - 0.44(\pm 0.03) \times \text{SVL}$, $F_{1,62} = 129$, $P < 0.001$, $r^2 = 0.67$; Santa Fe females: shrinkage, $164(\pm 16) - 0.65(\pm 0.06) \times \text{SVL}$, $F_{1,43} = 123$, $P < 0.001$, $r^2 = 0.74$; Genovesa males: shrinkage, $127(\pm 11) - 0.54(\pm 0.05) \times \text{SVL}$, $F_{1,26} = 119$, $P < 0.001$, $r^2 = 0.82$; Genovesa females: shrinkage, $85(\pm 20) - 0.44(\pm 0.10) \times \text{SVL}$, $F_{1,4} = 18$, $P = 0.01$, $r^2 = 0.82$. **b**, Relation between change in body length and survival time of adult iguanas on Santa Fe. Large adult individuals (SVL > 300 mm) that shrank most during the 1992–93 El Niño event survived longer (analysis of variance, $F = 21$, $P < 0.001$). Using other categories or relative shrinkage data exaggerates the observed trends². Data show mean survival time ($\pm$ 95% confidence intervals) after March 1993 for all individuals ($n=610$) measured from January to March 1992 and remeasured from January to March 1993.

Martin Wikelski*, Corinna Thom†

*University of Illinois at Urbana-Champaign, Department of Ecology, Ethology and Evolution, 515 Morrill Hall, 505 S. Goodwin Avenue, Urbana, Illinois 61801, USA; Max-Planck Institute for Behavioural Physiology, 82319 Seewiesen, Germany; and University of Bielefeld, VHF 33501, Germany
e-mail: wikelski@life.uiuc.edu
†Department of Biology, University of Würzburg, 97072 Würzburg, Germany

1. Andrews, R. M. in *Biology of the Reptilia* Vol. 13 (eds Gans, C. & Pough, F. H.) 273–320 (Academic, London, 1982).
2. Wikelski, M. & Trillmich, F. *Evolution* **51**, 922–936 (1997).

3. Wikelski, M., Carrillo, V. & Trillmich, F. *Ecology* **78**, 2204–2217 (1997).
4. Wikelski, M. & Trillmich, F. *Behaviour* **128**, 255–279 (1994).
5. Wolter, K. & Timlin, M. S. *Weather* **53**, 315–324 (1998).
6. Laurie, W. A. in *Global Ecological Consequences of the 1982–83 El Niño Southern Oscillation* (ed. Glynn, P.) 121–141 (Elsevier, Amsterdam, 1989).
7. Grant, P. R. *Ecology and Evolution of Darwin's Finches* (Princeton Univ. Press, New Jersey, 1986).
8. Cooper, J. E. & Laurie, W. A. *J. Comp. Pathol.* **97**, 129–136 (1987).

9. Laurie, W. A. & Brown, D. J. *Anim. Ecol.* **59**, 545–568 (1990).
10. Jarman, P. J. & Sinclair, A. R. E. in *Serengeti: Dynamics of an Ecosystem* (eds Sinclair, A. R. E. & Norton-Griffith, M.) 130–163 (Univ. Chicago Press, 1979).
11. Enlow, D. H. in *Biology of the Reptilia Vol. I* (eds Gans, C. et al.) 45–80 (Academic, London, 1969).
12. Grigoriev, A. I. et al. *Aviakosm. Ekolog. Med.* **32**, 21–25 (1998).
13. Strollo, F. et al. *Aviat. Space. Environ. Med.* **69**, 777–780 (1998).
14. Wronski, T. J. & Morey, E. R. *Metab. Bone Dis. Relat. Res.* **4**, 69–75 (1982).
15. Mundy, G. R. *Nature* **367**, 216–217 (1994).

Conservation

Biogeography of the Indonesian coelacanths

Living coelacanths (*Latimeria chalumnae*) are normally found only in the western Indian Ocean, where they inhabit submarine caves in the Comores Islands¹. Two specimens have since been caught off the island of Manado Tua, north Sulawesi, Indonesia, some 10,000 kilometres away². We sought to determine the ecological and geographic distribution of Indonesian coelacanth populations with a view to drawing up conservation measures for this extremely rare fish^{2,3}. During our explorations, we discovered two living Indonesian coelacanths 360 km southwest of Manado Tua.

Analysis of mitochondrial DNA from a single Indonesian specimen, described as a distinct species, *L. menadoensis*⁴, revealed significant differences with the published mitochondrial genome of *L. chalumnae*. These differences indicate that the Indonesian and Comoran coelacanths diverged 1.8–11.0 million years ago⁵.

In an attempt to find more fish, we performed a total of 34 dives in the research submersible *JAGO* down to a depth of 400 metres at points along the northern coast of Sulawesi, the Sangihe islands and the Bay of Tomini in central Sulawesi. We failed to find any coelacanths during seven dives off Manado Tua, where the two previously reported coelacanths were caught², but 360 km southwest of Manado Tua we found two coelacanths approximately 120 and 140 cm long in a deep carbonate cave at a depth of 155 m (water temperature, 17.8–20.1 °C).

The substrate and oceanographic conditions in the Indonesian and Comoran dive sites are completely different. The Comores have steep, young volcanic slopes with numerous lava caves, whereas the slopes of the Indonesian dive sites are older, more eroded and less steep, with very few caves, and these are mainly carbonate in origin. The Indonesian sites are exposed to strong currents (with an estimated peak velocity of 3–4 knots) of variable directions and with sudden up- and downwellings. In contrast, the Comoran habitat is frequently devoid of currents, and these are slower than 1 knot.

These differences may be important, as

Comoran coelacanths are sluggish, nocturnal drift hunters that feed on other fish and have a low metabolic rate⁶. During the day, they retreat to the still water of deep lava caves¹. The Indonesian sites would therefore seem to be unsuitable for sustaining a viable population of *L. chalumnae*.

So far, four Indonesian coelacanths have been found in a relatively small area along the north Sulawesi coast. We suggest that the population is very small and requires strict conservation measures. But we cannot exclude the possibility that the coelacanths of north Sulawesi are derived from a different area and drifted there with oceanic currents. The dominant current driving Indonesian throughflow water in the area of north Sulawesi is the southerly Mindanao current⁷, indicating that the potential source population may be in the southern Philippines or remote Pacific islands.

Furthermore, geological evidence and our genetic studies of the Comoran population indicate a young age for this population of less than 100,000 years. The reported older genetic separation of the Indonesian coelacanth of more than 1 million years would suggest that Comoran *L. chalumnae* probably derived from a third, as yet unknown, population.

The biogeography of the new coelacanth population remains enigmatic, although perhaps this is for the best. An undiscovered home is probably the best possible protection for these endangered fish.

H. Fricke*, K. Hissmann*, J. Schauer*, M. Erdmann†, M. K. Moosa‡, R. Plante§

*Max-Planck-Institut für Verhaltensphysiologie, 82319 Seewiesen, Germany

e-mail: fricke@mpi-seewiesen.mpg.de

†Department of Integrative Biology, University of California, Berkeley, California 94720, USA

‡Indonesian Institute of Science (LIPI), Ancol Timur, Jakarta 11048, Indonesia

§Centre d'Océanologie de Marseille, Station Marine d'Endoume, 13007 Marseille, France

1. Fricke, H. & Hissmann, K. *Mar. Biol.* **120**, 171–180 (1994).
2. Erdmann, M. V., Caldwell, R. L. & Moosa, M. K. *Nature* **395**, 335 (1998).
3. Forey, P. L. *Nature* **395**, 319–320 (1998).
4. Pouyaud, L. et al. *C. R. Acad. Sci.* **322**, 261–267 (1999).
5. Holder, M. T. et al. *Proc. Natl. Acad. Sci. USA* **96**, 12616–12620 (1999).
6. Fricke, H. & Hissmann, K. *Mar. Biol.* (in the press).
7. Gordon, A. L. & Fine, R. A. *Nature* **379**, 146–149 (1996).

Neuroscience

Extrapolating movement without retinal motion

In contrast to the perception of a stationary object that is briefly flashed in the dark, a continuously visible moving object is seen as being ahead of its actual position at the time of the flash. An explanation for this simple effect, in which a stimulus moving on the retina is seen as being further along its path and not where it was in space when its signal impinged on the retina, is keenly debated^{1–6}. We show here that this illusion is not just limited to retinal motion, and that perceptual mislocalization occurs even when stimulus motion is inferred entirely from extra-retinal information, for example by movement of the observer's head or whole body, without retinal motion. The phenomenon may therefore rely on a much more general mechanism.

Nijhawan^{1,2} originally suggested that the illusion is the product of a brain process that tries to overcome at least some of the visual transmission delay (more than 50 ms) through extrapolation, so that we should be able, for instance, to catch a moving object accurately. Other observations^{3,4} indicated that the apparent stimulus misalignment could be the passive consequence of the difference in the afferent delay between a stimulus that is moving on the retina and a stationary one. The physiological properties of the retina itself may even be sufficient to produce the misalignment⁶. These explanations are all based on retinal motion, however, which is the only condition under which the phenomenon has so far been demonstrated.

We set up two situations in which the movement of the stimulus in space was generated entirely by the movement of the

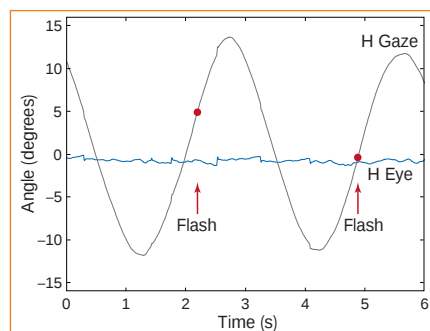


Figure 1 Recording of horizontal eye movement relative to external space, and eye movement relative to the head. The subject rotated his head sinusoidally back and forth while fixating a light-emitting diode array mounted on a recording Eyelink helmet. Flashes occurred at times indicated by arrows and were perceived as lagging behind the continuously lit stimulus. Subjects estimated the perceptual misalignment in this condition by comparing it afterwards, head fixed, with the misalignment of the same stimulus rotating on a turntable at 30 revolutions per minute, as previously^{1,2}.

observers, without retinal motion. In the first situation, five subjects (two authors and three naive subjects) wearing the helmet of an eye-tracking system (SensoMotoric Eyelink) focused their gaze on a vertical bar made up of light-emitting diodes (LEDs). The bar was mounted on the helmet 36 cm in front of the eyes. In complete darkness, subjects were instructed to rotate their heads horizontally back and forth (at 20–30 degrees, with a rhythm of about 0.3–0.4 Hz). The lower two-thirds of the bar (26 mm × 1.5 mm) was continuously lit, while the upper one-third of the bar was flashed for 6 ms halfway through the head movements.

The recordings showed that, while gaze (the position of the eye in space) moved approximately 25 degrees per cycle (Fig. 1), the total eye displacements relative to the head were less than 1 degree per cycle. Furthermore, no eye displacement in the head occurred at the time of the flashes, so the motion of the stimulus on the retina was minimal compared with the motion of the stimulus in space.

When the subjects were asked to judge the position of the array with respect to the position of the flashed LEDs, however, they invariably reported that the continuously lit LEDs were ahead of the flashed ones. Even though all subjects, including the naive ones, had seen the LED array before the experiment and knew that all the LEDs were fixed on a single rigid bar, they still perceived a misalignment of several degrees of visual angle between the continuously lit and the flashing segments, as in earlier studies in which there was retinal motion^{1,2}. When the subjects were continuously rotated in a chair at 20 revolutions per minute, the visual stimulus was the same. The continuously lit moving stimulus was seen as leading, most clearly at the start of rotation, and as lagging during the final deceleration.

The brain has no direct access to the timing of external events because input delays are variable and therefore unreliable³. When moving stimuli are involved, uncertainty about the time of an event (for example, of a flash) can translate into uncertainty about position. The time–position ambiguity resulting from processing delays may affect the perception of stimulus motion, regardless of how the brain has access to this information. Whether the cue derives from retinal, oculomotor, vestibular or proprioceptive signals, the perceived position of a moving object may be extrapolated in the same way.

John Schlag, Rick H. Cai, Andrews Dorfman, Ali Mohempour, Madeleine Schlag-Rey

Department of Neurobiology, School of Medicine, UCLA, Los Angeles, California 90095, USA

e-mail: jschlag@ucla.edu

1. Nijhawan, R. *Nature* **370**, 256–257 (1994).
2. Nijhawan, R. *Nature* **386**, 66–69 (1997).
3. Purushothaman, G., Patel, S. S., Bedell, H. E. & Ogmen, H. *Nature* **396**, 424 (1998).
4. Whitney, D. & Murakami, I. *Nature Neurosci.* **1**, 656–657 (1998).
5. Gegenfurtner, K. *Nature* **398**, 291–292 (1999).
6. Berry, M. J., Brivanlou, I. H., Jordan, T. A. & Meister, M. *Nature* **398**, 334–338 (1999).

Cognition

Numerical memory span in a chimpanzee

A female chimpanzee called Ai has learned to use Arabic numerals to represent numbers¹. She can count from zero to nine items, which she demonstrates by touching the appropriate number on a touch-sensitive monitor^{2,3}, and she can order the numbers from zero to nine in sequence^{4–6}. Here we investigate Ai's memory span by testing her skill in these numerical tasks, and find that she can remember the correct sequence of any five numbers selected from the range zero to nine.

Humans can easily memorize strings of codes such as phone numbers and postcodes if they consist of up to seven items, but above this number they find it much harder. This 'magic number 7' effect, as it is known in human information processing⁷, represents a limit for the number of items that can be handled simultaneously by the brain.

To determine the equivalent 'magic number' in a chimpanzee, we presented our subject with a set of numbers on a screen, say 1, 3, 4, 6 and 9. She had already displayed close to perfect accuracy when required to choose numerals in ascending order, but for this experiment all the remaining numbers were masked by white squares once she had selected the first number. This meant that, in order to be correct in a trial, she had to memorize all the numbers, as well as their respective positions, before making the first response. Chance levels with three, four and five items were 50, 13 and 6%, respectively.

Ai scored more than 90% with four items and about 65% with five items, significantly above chance in each case. In normal background trials, response latency was longest for the first numeral and much

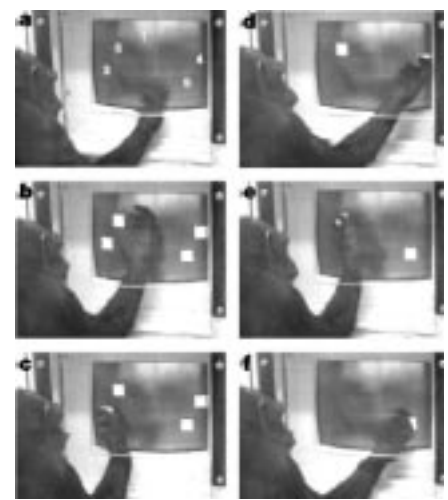


Figure 1 The chimpanzee Ai performing the five-number ordering task in the 'masking' trial. Five numbers (1, 3, 4, 6 and 9) are presented on the touch-sensitive monitor. **a, b**, Ai correctly chooses the number 1 as the lowest of the series (**a**), at which point the remaining numbers are automatically masked (**b**). **c–f**, She continues to identify the numbers one by one in ascending order (**c–e**), ending with the 9 (**f**). See Supplementary Information and <http://www.pri.kyoto-u.ac.jp> for more details.

shorter for all the others, indicating that Ai inspected the numbers and their locations and planned her actions before making her first choice. In masking trials, response latency increased only for the choice directly after the onset of masking, but this latency was similar to those recorded in background trials, indicating that successful performance did not depend on spending more time memorizing the numbers.

In one testing session, after Ai had chosen the correct number and all the remaining items were masked by white squares, a fight broke out among a group of chimpanzees outside the room, accompanied by loud screaming. Ai abandoned her task and paid attention to the fight for about 20 seconds, after which she returned to the screen and completed the trial without error.

Ai's performance shows that chimpanzees can remember the sequence of at least five numbers, the same as (or even more than) preschool children. Our study and others^{8–10} demonstrate the rudimentary form of numerical competence in non-human primates.

Nobuyuki Kawai, Tetsuro Matsuzawa

Primate Research Institute, Kyoto University,

Table 1 Performance in masking trials

Type of trial	Numbers	Trials	Number (% correct)						Response time (ms)				
			1st	2nd	3rd	4th	5th	Total	1st	2nd	3rd	4th	5th
Normal	2	405	98	100	—	—	—	98	676	420	—	—	—
Normal	3	433	97	97	100	—	—	94	710	424	420	—	—
Normal	4	451	93	96	98	100	—	87	754	439	407	412	—
Normal	5	421	90	93	94	99	100	78	799	448	415	430	408
Masking	3	200	98	91	100	—	—	89	768	533	465	—	—
Masking	4	20	100	100	95	100	—	95	717	390	432	437	—
Masking	5	20	95	95	89	81	100	65	721	446	426	466	411

1st, 2nd, 3rd, 4th and 5th refer to the numbers in a sequence in ascending order.

Inuyama, Aichi 484-8506, Japan
e-mail: matsuzaw@pri.kyoto-u.ac.jp

1. Matsuzawa, T. *Nature* **315**, 57–59 (1985).
2. Matsuzawa, T., Itakura, S. & Tomonaga, M. in *Primate Today* (eds Ehara, A., Kumura, T., Takenaka, O. & Iwamoto, M.) 317–320 (Elsevier, Amsterdam, 1991).
3. Murofushi, K. *Jpn. Psychol. Res.* **39**, 140–153 (1997).
4. Tomonaga, M., Matsuzawa, T. & Itakura, S. *Primate Res.* **9**, 67–77 (1993).
5. Biro, D. & Matsuzawa, T. *J. Comp. Psychol.* **113**, 178–185 (1999).
6. Tomonaga, M. & Matsuzawa, T. *Anim. Cogn.* (in the press).
7. Miller, G. A. *Psychol. Rev.* **63**, 81–97 (1956).
8. Rumbaugh, D., Savage-Rumbaugh, E. S. & Hegel, M. J. *Exp. Psychol. Anim. Behav. Process.* **13**, 107–115 (1987).
9. Brannon, E. & Terrace, H. *Science* **282**, 746–749 (1998).
10. Boysen, S., Mukobi, K. & Berntson, G. *Anim. Learn. Behav.* **27**, 229–235 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Medical physics

Explaining the T-wave shape in the ECG

The heartbeat is recorded on an electrocardiogram (ECG) as a characteristic trace determined by changes in the electrical activity of the heart muscle. The T wave is a component of this waveform that is associated with the repolarization phase of the action potentials¹. It is asymmetrical in healthy subjects, but tends to become symmetrical with heart disease². The reason for the T-wave shape is not clear³. Here we show that T waves become more symmetrical as a result of an increase in the dispersion of the regional repolarization of cardiac muscle.

The exact sequence of depolarization and repolarization of the action potential (the potential difference that arises between the intracellular and extracellular fluid) in three-dimensional heart muscle is complex, but to model the electrocardiogram on the body surface, only the heart surface potentials need to be known⁴, assuming that the myocardium has uniform and isotropic conductivity. Although the repolarization

on any individual cell is timed to the initiation of its action potential, repolarization occurs smoothly and systematically across the epicardial surface^{5,6}. Repolarization of the surface of the left ventricle occurs first epicardially in the postero-basal region, then at the septal wall and apex⁶, and finally at the endocardial surface⁷.

We used this spatial sequence on a simple model of the left ventricle which allowed the body-surface 12-lead T waves of the ECG to be calculated with the standard repolarization phases of an action potential⁸ and standard modelling equations⁴. The body was represented by an elliptical cylinder⁹ and the heart by a truncated ellipsoid (axis diameters, 7.0 and 6.6 cm; height from base, 7.4 cm; displaced 4 cm left and 6 cm forward of the body axis, and rotated 25 degrees forward and 40 degrees to the left).

We evaluated this model twice: by using a fixed, normal shape for the action potential, and then by using three regions, each with a different shape for the action potential, with a 10% increase in model parameters⁸ at the endocardium and with a 10% decrease at the apex. These changes to the parameters generated abnormally shaped action potentials. To make sure that the results did not depend on the specific model values, we calculated the form of T waves for the initial heart position and for separate shifts of the axes of ± 2 cm and rotations of ± 10 degrees, to calculate error bars.

To represent differences in the regional dispersion of repolarization times, we varied the time delay between the first and last action potentials, and thus between the earliest and latest region to repolarize, from 10 to 150 ms in steps of 10 ms, which ranges from normal to abnormal dispersion, and linearly interpolated the time delays for intermediate regions. We calculated a symmetry ratio from the body-surface T waves as the ratio of the areas under the two sections of the T curves

(beginning-to-peak compared with peak-to-end).

Our results show that low dispersion is represented by asymmetrical T waves, and high dispersion by increasingly tall and symmetrical, clinically hyperacute T waves that tend to a symmetry ratio of unity (Fig. 1). This agrees with the accepted symmetry ratio for normal T waves of 1.5 (ref. 10) and with the expected relation between tall, symmetrical T waves and abnormal repolarization. Abnormal shapes in the action potential result in differences in the T-wave symmetry ratio, but still tend to produce symmetrical T waves for high dispersion.

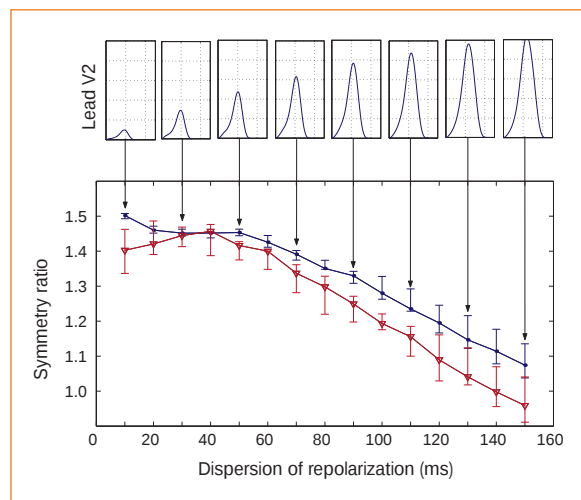
This finding can be explained by considering a simplified situation, based on the assumption that the heart contains only two regions, which give rise to action potentials that have identical shapes but are displaced in time. The resulting potential gradients, and ultimately the shape of the T wave, can be approximated to the difference between the two action potentials at each instant. When dispersion, or time displacement, is small, this difference will result in an asymmetrical waveform (as a consequence of the shape of the action-potential repolarization phase), and when it is large (with the second action potential beginning to repolarize after the first has mostly repolarized), a symmetrical waveform will tend to be produced.

Such a simple model can be expressed mathematically and computed easily but is not convincing without the calculations used here, with more realistic heart and torso geometries and gradual differences in the initiation of action-potential repolarization across the myocardium. Confidence in the model is increased by the resulting T-wave shapes in the 12 ECG leads.

Analysis of T-wave symmetry offers a new clinical indication of dispersion, which would be valuable for patients with high dispersion, who are more likely to die suddenly¹¹; the current methods used to measure dispersion are unsatisfactory and are prone to errors¹². Our results indicate that there is a link between T-wave symmetry and the abnormal regional dispersion of repolarization.

Diego di Bernardo, Alan Murray
Regional Medical Physics Department,
Newcastle University, Freeman Hospital,
Newcastle upon Tyne NE7 7DN, UK

Figure 1 T-wave symmetry ratios for a range of dispersions of repolarization for the two implementations of the model (blue circles, single action potential; red triangles, three regions with different APs) for the mean of precordial leads V2 to V6. V1 was omitted as it was sometimes biphasic. Lines are drawn through the symmetry ratios for the initial heart position; error bars give the maximum range calculated for the different heart positions. The mean difference between limb and precordial leads for the symmetry ratio was 0.026 for the model of the single action potential, and 0.005 for three action potentials. The graphs at the top show the T waves for the single normal action potential in an example lead V2; each grid square is 0.5 mV tall and 200 ms wide, as in standard ECG paper recordings.



1. Noble, D. & Cohen, I. R. A. *Cardiovas. Res.* **12**, 13–27 (1978).
2. Conover, M. B. *Understanding Electrocardiography: Arrhythmias and the 12-lead ECG* (Mosby Year Book, St Louis, Missouri, 1992).
3. Yan, G.-X. & Antzelevitch, C. *Circulation* **98**, 1928–1936 (1998).
4. Simms, H. D. & Geselowitz, D. B. *J. Cardiovas. Electrophysiol.* **6**, 522–531 (1995).
5. Franz, M. R., Bargheer, K., Rafflenbeul, W., Haverich, A. & Lichten, P. R. *Circulation* **75**, 379–386 (1987).
6. Cowan, J. C. et al. *Br. Heart J.* **60**, 424–433 (1988).
7. Higuchi, T. & Nakaya, Y. *Am. Heart J.* **108**, 290–295 (1984).
8. Wohlfart, B. *Eur. Heart J.* **8**, 409–416 (1987).
9. Gelernter, H. L. & Swihart, J. C. *Biophys. J.* **4**, 285–301 (1964).
10. Merri, M., Benhorin, J., Alberti, M., Locati, E. & Moss, A. J. *Circulation* **80**, 1301–1308 (1989).
11. Barr, C. S. et al. *Lancet* **343**, 327–329 (1994).
12. Murray, A. et al. *Br. Heart J.* **71**, 386–390 (1994).

The language of covalent histone modifications

Brian D. Strahl & C. David Allis

Department of Biochemistry and Molecular Genetics, University of Virginia Health Science Center, Charlottesville, Virginia 22908, USA

Histone proteins and the nucleosomes they form with DNA are the fundamental building blocks of eukaryotic chromatin. A diverse array of post-translational modifications that often occur on tail domains of these proteins has been well documented. Although the function of these highly conserved modifications has remained elusive, converging biochemical and genetic evidence suggests functions in several chromatin-based processes. We propose that distinct histone modifications, on one or more tails, act sequentially or in combination to form a 'histone code' that is, read by other proteins to bring about distinct downstream events.

How eukaryotic genomes are manipulated within a chromatin environment is a fundamental issue in biology. At the heart of chromatin structure are highly conserved histone proteins (H3, H4, H2A, H2B and H1) that function as building blocks to package eukaryotic DNA into repeating nucleosomal units that are folded into higher-order chromatin fibres^{1,2} (Fig. 1). Once thought of as static, non-participating structural elements, it is now clear that histones are integral and dynamic components of the machinery responsible for regulating gene transcription. The same is probably true for other DNA-templated processes such as replication, repair, recombination and chromosome segregation.

An extensive literature documents an elaborate collection of post-translational modifications including acetylation, phosphorylation, methylation, ubiquitination and ADP-ribosylation³ that take place on the 'tail' domains of histones. These tails, which protrude from the surface of the chromatin polymer and are protease sensitive, comprise ~25–30% of the mass of individual histones^{3,4}, thus providing an exposed surface for potential interactions with other proteins (for example, Sir3/4 and Tup1 proteins in yeast^{5,6}). Because of the inherent disordered nature of histone tails, their precise location in higher-order fibres and the atomic details of their structure are not known^{1,7}.

Long-standing models have suggested that histone modifications may alter chromatin structure by influencing histone–DNA and histone–histone contacts^{4,8}. However, growing awareness of the remarkable diversity and biological specificity associated with distinct patterns of covalent histone marks has caused us and others^{9–13} to favour the view that a histone 'language' may be encoded on these tail domains that is read by other proteins or protein modules. We refer to this language as the 'histone code' and present evidence supporting the existence of this language and discuss some potential ramifications. To illustrate the potential complexity of covalent marks decorating a single histone tail, we have chosen to focus most of our discussion on a short stretch of core histone H3. However, many of the concepts presented here are likely to apply to all of the histone termini and, in particular, to that of histone H4^{10,14}.

Lysine acetylation sets the stage

Of the modifications listed above, histone acetylation has been the most studied and appreciated¹⁴. Fuelled, in part, by the discovery of enzymes responsible for bringing about the steady-state balance of this modification—histone acetyltransferases (HATs) and histone deacetylase (HDACs)—compelling evidence has recently been provided that acetylation of specific lysine residues in the amino termini of the core histones plays a fundamental role in transcriptional regulation^{2,15}.

In H3 from most species, the main acetylation sites include lysines 9, 14, 18 and 23 (refs 3, 16), and, as is the case with the functionally redundant tail of its nucleosomal partner, H4¹⁴, selected lysines become acetylated during specific cellular processes (Figs 1b and 2).

Transcription-linked acetylation, catalysed by the GCN5 family of HATs, shows a preference for lysine 14 of H3 *in vitro*¹⁷ although an expanded set of lysine residues is likely to be used *in vivo*^{18,19}. How is this acetylation site specificity in H3 brought about?

Solution and crystal structure data of various members of the GCN5 HAT family, including co-crystals of the enzyme with H3 tail peptides²⁰, have begun to yield important insights into the enzymatic mechanisms underlying the site specificity of these HATs^{21–25}. One important concept to emerge from these studies is that residues outside the preferred lysine 14 acetylation site in H3 are important for histone-binding specificity. For example, glycine 13 and proline 16 have a critical role in leading to a restricted GCN5–H3 peptide recognition site, G-K14*-X-P (ref. 20). Thus, as is the case with protein kinases and phosphatases, short preferred consensus motifs are likely to exist for individual HATs and HDACs²⁶ which help to establish the final histone code.

Acetylation of specific lysine residues in H3 is also associated with biological processes apart from transcription (Fig. 2). During DNA replication, for example, new histones are rapidly synthesized and assembled onto the replicated DNA. H3 and H4 are brought to replicating chromatin in a pre-acetylated state that becomes erased after replication is completed and the newly assembled chromatin matures^{27,28}. Whereas the sites of deposition-related H4 acetylation are highly conserved^{29,30} (for example, lysines 5 and 12; see Fig. 2), the situation with H3 is less clear. However, lysine 9 in H3 appears to have a more dominant role in histone deposition and chromatin assembly in some organisms^{17,27,30}. The finding that a chromatin assembly complex in *Drosophila*, called RCAF (for replication-coupling assembly factor), contains H4 specifically acetylated at lysines 5 and 12 suggests that these acetylation sites play an important role in chromatin assembly³¹. Does this acetylation pattern represent a code that has been deciphered by a component of a histone chaperone complex?

Finally, we note that the spacing between acetyltable lysines (Fig. 1b) is strikingly regular in the amino termini of many histones (for example, lysines at 9, 14, 18 and 23 in H3; and 5, 8, 12 and 16 in H4), and, curiously, this spacing periodicity is reminiscent of that of an α -helix (that is, 3.6 residues). To our knowledge, no group has systematically attempted to expand or contract the characteristic three-to-four residue spacing between many known acetylation sites. Along this line, the alternating, and seemingly, invariant pattern of deposition-related acetylation wherein lysines 5 and 12, but not lysines at 8 and 16, are acetylated in newly synthesized H4 is particularly intriguing^{29,30} (Fig. 2).

Beyond histone acetylation

Phosphorylation, particularly that of histones H1 and H3, has long been implicated in chromosome condensation during mitosis^{32,33}. However, converging evidence suggests that H3 phosphorylation

(specifically serine 10; see Fig. 1b) is also directly correlated with the induction of immediate-early genes such as *c-jun*, *c-fos* and *c-myc*^{34–36}. Mutations in Rsk-2, recently shown to be an H3 kinase *in vitro*, are associated with Coffin–Lowry syndrome in humans and result in a loss of epidermal growth factor-stimulated H3 phosphorylation *in vivo*^{9,37}. Transcriptional activation in response to mitogenic and other stimuli are altered in Coffin–Lowry cells³⁸, suggesting a potential direct role for H3 phosphorylation in regulating gene transcription through a remodelling step that is most consistent with chromatin decondensation, a result seemingly at odds with the use of this mark in chromosome condensation.

The potential importance of the serine 10 phosphorylation mark in H3 is strengthened by the finding that MSK1, a kinase activated by growth factor and stress stimuli, also phosphorylates H3 *in vitro*³⁹. Interestingly, another H3 kinase has recently been identified that is associated with dosage compensation in flies⁴⁰. In *Drosophila*, equalization of transcription from the sex chromosomes is achieved by a twofold upregulation of transcription from the male X chromosome⁴¹ that is associated with acetylation of H4 at lysine 16⁴². Thus, H4 acetylation on lysine 16, possibly in concert with H3 phosphorylation at serine 10, may establish a combinatorial mark that leads to enhanced transcription from the male X chromosome.

What about other histone modifications?

Just as histone acetylation and phosphorylation have become topics

of renewed interest, we expect other known histone modifications will soon be back in the limelight. Methylation of lysine and/or arginine residues, for example, are among the least understood post-translational modifications affecting histones. This is partly because the responsible enzymes are not known, immunological reagents selective to detect histone methylation do not exist, and, unlike histone acetylation and phosphorylation, the charge on the individual lysine and arginine residues is not greatly affected, making electrophoretic resolution of methylated histones difficult. H3 and H4 are the predominant histones modified by methylation, and sequencing studies from many organisms indicate that multiple lysines in H3 (4, 9 and 27) are the preferred sites of methylation^{3,43,44}. Moreover, lysine residues can be mono- di- or trimethylated, adding yet another layer of complexity to this histone mark.

The recent discovery of a nuclear receptor co-activator-interacting protein, called CARM1, which possesses arginine-specific, histone H3-selective methyltransferase (HMT) activity⁴⁵ provides evidence to support the notion that histone methylation contributes to transcriptional activation. CARM1 HMT activity is required for ligand-dependent transcriptional activation, and the finding that CARM1 functions through association with co-activators containing HAT activity is particularly intriguing given that Rsk-2 interacts with the transcriptional co-activator CREB-binding protein (CBP)⁴⁶. These data suggest that large multisubunit enzyme complexes containing multiple histone- and non-histone-modifying activities^{9,47} work in concert with other chromatin remodelling machines^{48,49} to regulate gene transcription (Fig. 3, left panel).

The 'histone code' hypothesis

Considering only the electrostatic requirements for folding the chromatin polymer^{4,8,50}, histone acetylation, through the neutralization of positive charge, and histone phosphorylation, through the addition of negative charge, would probably cause decondensation of the chromatin fibre⁵¹. Thus, the use of multiple marks on histone tails (that is, combining acetylation and phosphorylation) could serve to amplify the readout of upstream signalling pathways causing greater changes in the overall charge density of tails that

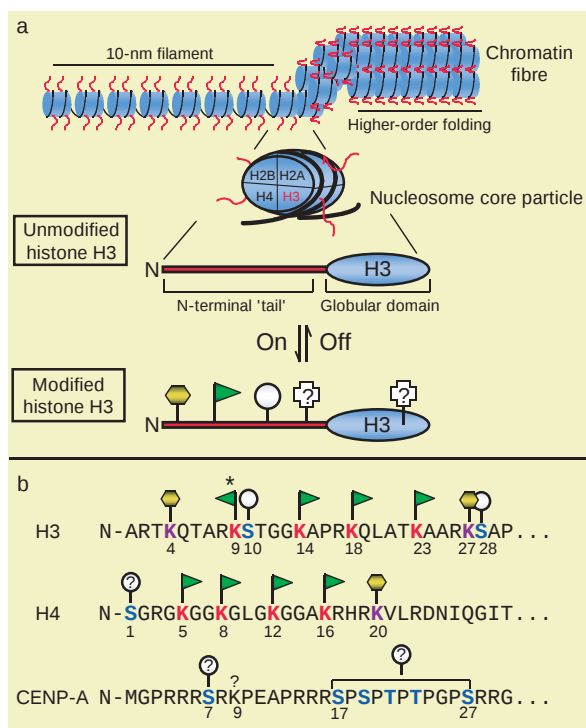


Figure 1 Chromatin organization and the tail of histone H3. **a**, General chromatin organization. Like other histone 'tails', the N terminus of H3 (red) represents a highly conserved domain that is likely to be exposed or extend outwards from the chromatin fibre. A number of distinct post-translational modifications are known to occur at the N terminus of H3 including acetylation (green flag), phosphorylation (grey circle) and methylation (yellow hexagon). Other modifications are known and may also occur in the globular domain. **b**, The N terminus of human H3 is shown in single-letter amino-acid code. For comparison, the N termini of human CENP-A, a centromere-specific H3 variant, and human H4, the nucleosomal partner to H3, are shown. Note the regular spacing of acetylable lysines (red), and potential phosphorylation (blue) and methylation (purple) sites. The asterisk indicates the lysine residue in H3 that is known to be targeted for acetylation as well as for methylation; lysine 9 in CENP-A (bold) may also be chemically modified (see text). The above depictions of chromatin structure and H3 are schematic; no attempt has been made to accurately portray these structures.

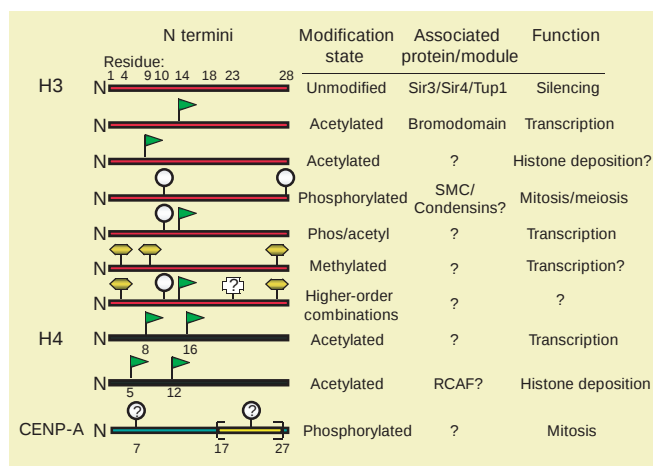


Figure 2 The 'histone code' hypothesis. Histone modifications occur at selected residues and some of the patterns shown have been closely linked to a biological event (for example, acetylation and transcription). Emerging evidence suggests that distinct H3 (red) and H4 (black) tail modifications act sequentially or in combination to regulate unique biological outcomes. How this hierarchy of multiple modifications extends (depicted as 'higher-order combinations') or how distinct combinatorial sets are established or maintained in localized regions of the chromatin fibre is not known. Relevant proteins or protein domains that are known to interact or associate with distinct modifications are indicated. The CENP-A tail domain (blue) might also be subjected to mitosis-related marks such as phosphorylation; the yellow bracket depicts a motif in which serines and threonines alternate with proline residues.

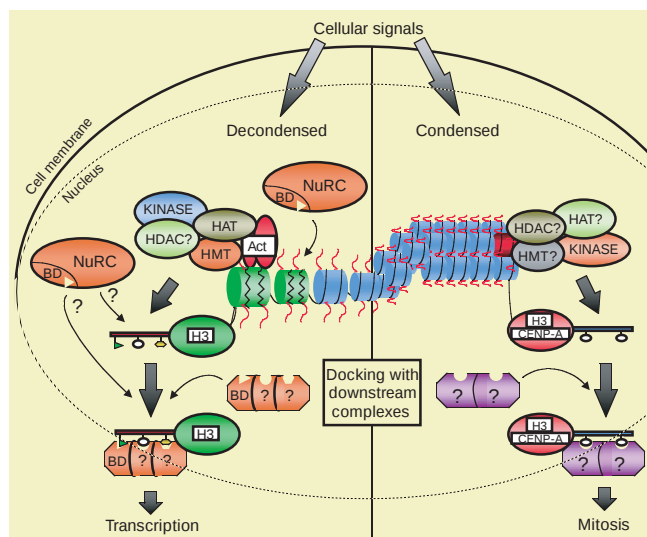


Figure 3 Coordinated recruitment of histone-modifying activities. Recent discoveries suggest that distinct histone-modifying activities interact to form multisubunit complexes that probably work in concert with nucleosome remodelling complexes (NuRCs; for example, Swi/Snf, RSC, NURF) to remodel chromatin. Interactions demonstrated thus far include CARM1, a histone methyltransferase (HMT), with histone acetyltransferase (HAT)-containing coactivators that interact with nuclear receptors (Act)⁴⁵ and Rsk kinase with the CBP/p300 HAT⁴⁶. Not depicted is the possibility of non-histone substrates being *bona fide* targets of these activities. In addition to remodelling nucleosomes (indicated by zigzag DNA), NuRCs may chemically modify and/or bind histone tails. Binding of a NuRC or HAT

complex (depicted as single ovals) to histone tails may be mediated by the bromodomain (BD). Although many of the complexes identified to date are implicated in events leading to transcription (left panel; green nucleosomes), we speculate that similar, but unique, complexes may exist that modify and direct chromatin condensation (right panel; red nucleosome). This 'code' of modifications may dictate the biological outcome through changes caused in higher-order chromatin structure or may direct the downstream biological effect by recruiting and interacting with docking proteins or complexes that remain to be identified. Although H3 and CENP-A are the only histones depicted here, it is likely that all other histones are subjected to this type of regulation.

lead to greater changes in the chromatin structure of target genes. Indeed, evidence supports synergism between histone acetylation and phosphorylation in the induction of immediate-early genes after mitogenic stimulation⁵², and it seems likely that these signalling pathways may converge at other loci as well³⁵.

H3 phosphorylation at serine 10, possibly in conjunction with phosphorylation at serine 28 (Figs 1b and 2; and see below), is also required for proper segregation and condensation of chromosomes during mitosis and meiosis^{53,54}. If the function of H3 phosphorylation is to 'open' chromatin, how then can H3 phosphorylation at the same site also be involved in chromosome condensation? This question seems to beg a simple answer: perhaps a single histone modification does not function alone. We will refer to the hypothesis—that multiple histone modifications, acting in a combinatorial or sequential fashion on one or multiple histone tails, specify unique downstream functions—as the histone code hypothesis. What is the evidence in support of this hypothesis?

Serine 28 in H3 is embedded in surrounding sequences similar to serine 10 (that is, both are R-K-S*), and data has shown that serine 28 is phosphorylated during chromosome condensation in mammalian cells⁵⁵ (Fig. 1b). Whether serine 28 is phosphorylated during interphase or during immediate-early gene induction is not yet known. Thus, the formal possibility remains that multiple phosphorylation events on the same histone tail, or on several tails, may be required for efficient chromosome condensation during mitosis and meiosis (see below).

Even if the H3 tail is doubly marked by serine 10 and serine 28 phosphorylation during mitosis, it seems likely that this is not the complete story regarding chromosome condensation. H3 phosphorylation at serine 10 initiates in the pericentric heterochromatin, an A/T-rich region of satellite DNA closely associated with centromeric DNA⁵³. Centromeric DNA itself is packaged with specialized proteins, one of which is a specialized H3 'variant' found in both yeast and humans, CENP-A⁵⁶. CENP-A differs from H3 primarily in its unique N-terminal tail (Fig. 1b). Apart from lysine 9, arginine residues comprise all other positively charged side chains. Thus, the role of acetylation at lysine 9 (if it occurs) probably differs from that

of acetylation of canonical H3 in association with transcription. Moreover, the CENP-A tail contains many serine and threonine residues, which raises the possibility that this specialized H3, like the main H3, becomes phosphorylated during mitosis. An alternating S/T/G-P motif repeats five times in this tail generating an 11-amino-acid stretch flanked on either side by arginine residues (Fig. 1b). This motif suggests that CENP-A may be phosphorylated during mitosis, and it will be interesting to determine whether the CENP-A and H3 proteins are substrates for the same or unique sets of kinases and phosphatases. It would also be of interest to determine whether the SMC/condensing proteins, which have a central role in mitotic chromosome condensation, bind to these tails and contain histone-modifying activities⁵⁷ (Fig. 2).

The enzymology of multiple histone modifications

The existence of multiple modifications within a short stretch of the same histone tail (Fig. 1b), begs the question: how is a complex, multimark code established and maintained in the first place? One attractive hypothesis is that covalent modification of a histone tail by one enzyme influences the rate or efficiency with which a second enzyme follows using the now-modified histone tail as substrate. Does site-specific phosphorylation or methylation influence the ability of a HAT to recognize and bind to the tail? Alternatively, does a histone kinase or methylase care whether a histone tail is acetylated at a specific lysine residue? To that end, we point out that many modifications are close enough to each other on the histone tail (Fig. 1b) to influence, positively or negatively, the ability of enzymes to further modify these residues. Along this line, modifications on one histone tail might influence the outcome of other enzymatic activities acting on other histone tails.

The explosion of recent discoveries of histone-modifying enzymes, many available in recombinant form, paves the way for future experimental tests of some of these questions. Structural studies with modified histone substrates will be necessary to determine which residues, if any, are used to stabilize or promote interactions with modified substrates. Site-directed mutagenesis of these residues, followed by *in vivo* and *in vitro* assays, will help to

dissect meaningful functional relationships.

How is the histone code read?

Could histone modifications exist simply to regulate chromatin structure? Core histone acetylation alone has been shown to relax higher-order chromatin structure *in vitro*, and to promote factor-binding to cognate DNA elements^{4,8}. Alternatively, histone modifications could also act as specific 'receptors' to recruit unique biological complexes that mediate downstream function (Figs 2 and 3). Phosphorylation of H3 at serine 10 is closely associated with both chromosome condensation during mitosis and immediately early gene induction following mitogenic stimulation. One possible explanation for this discrepancy is shown in Fig. 3. Here we envisage that a phosphorylation mark alone, or in combination with other marks (such as phosphorylation at serine 28), may recruit a binding factor that, in turn, has a role in mediating chromosome condensation and segregation. In contrast, a distinct mark or set of marks (for example, phosphorylation and acetylation at residues 10 and 14, respectively) may provide a unique binding surface to recruit factors promoting decondensation and transcription (for example, Swi/Snf; see below). Assuming that phosphorylation of CENP-A occurs during mitosis, this marked tail may provide an attractive binding surface for a kinase that carries out general H3 phosphorylation at serine 10 and/or serine 28. In the specific case of the CENP-A tail (Fig. 1b), clusters of arginine residues are in some cases interrupted by serine residues (for example, R-R-S*-R-K) where the marked serine is serine 7. Phosphorylation at these positions may possibly serve to modify interactions with proteins that recognize this basic patch in the CENP-A tail.

One appealing feature of the histone code hypothesis is that it offers a possible explanation for 'exceptions' to the general rule that histone acetylation correlates positively with gene activation, whereas histone deacetylation acts to create repressive chromatin. For example, recent studies on the mouse mammary tumour virus (MMTV) promoter suggest that histone acetylation is actually involved in transcriptional repression (T. K. Archer and C. L. Smith, personal communication). Similarly, mutations in the HDAC homologue RPD3 cause enhancement of position effect variegation in flies, not suppression as would first be expected⁵⁸. Along these lines, pericentric heterochromatin in flies⁴² and silent loci in yeast⁵⁹ are marked by acetylation of lysine 12 of H4.

The disparity of having histone acetylation linked to both gene activation and repression is reminiscent of the situation with histone H3 phosphorylation being linked to both chromosome condensation and immediate-gene induction. Part of the solution to this paradox may be in having unique histone codes read by distinct sets of proteins that then bring about different downstream responses. If correct, it may be that mitosis-specific HATs, HDACs and HMTs act during chromosome condensation and that distinct sets of histone-modifying enzymes mark chromatin for decondensation during gene activation (Fig. 3).

Who reads the code?

Direct evidence that H3 and H4 tails can act as specific 'receptors' has been provided for Sir3 and Sir4⁵ and Tup1/Ssn6 (ref. 6) proteins involved in transcriptional silencing and repression in yeast. Moreover, Tup1 binding is influenced by the acetylation state of the H3 and H4 N-terminal tails: unacetylated or monoacetylated H3 and H4 are more strongly bound by Tup1 compared with hyperacetylated H3 or H4, suggesting that alterations in tail structure and/or charge due to acetylation can modulate non-histone protein/tail binding interactions. Whether or not other histone marks (such as phosphorylation, methylation) regulate these interactions further remains an important issue for future studies.

Recent evidence shows that the bromodomain of human PCAF (P300/CBP-associated factor), a domain of little known function which is shared between many, but not all HATs, binds acetylated

lysine in the context of H3 and H4 tail sequences⁶⁰. This result suggests that protein motifs may have evolved to recognize histone modifications⁶¹ (Fig. 3). Precedence for this type of receptor–ligand interaction already exists in nature. Phosphorylated tyrosine, for example, is known to be read in specific contexts by SH2-containing modules that, in turn, have an impact on downstream biological events⁶². Have similar mechanisms evolved for lysine acetylation, as well as other covalent histone modifications⁶¹?

Relevant to this discussion may be the observation that the spacing of acetylable lysines in the N termini of many histones is regular. Is this spacing part of the histone recognition motif? To that end, we note that many bromodomain-containing proteins have two adjacent bromodomain modules (for example, TAF_{II}250). Whether in these proteins each bromodomain functions independently or synergistically to bind acetylated lysines, in one or more histone tails, is not known, but remains an intriguing possibility.

If bromodomains bind acetylated histones, then this suggests that other chromatin-associated polypeptides containing this domain may also function through recognition and binding to specific acetylation patterns on histones (Fig. 3). For example, components of the Swi/Snf and RSC (for remodels the structure of chromatin) family of ATP-dependent chromatin remodelling enzymes contain bromodomains⁶¹. Is it possible that these remodelling complexes function through a combination of factor recruitment⁴⁸ and recognition of distinct acetylation patterns on the histone tails at promoters regulated by these complexes⁴⁹? Some support for this is provided by genetic evidence that suggests that Swi/Snf function is partially redundant with the functions of Gcn5-containing HAT complexes at specific promoters^{63–65} and by studies that show that ATP-dependent remodelling complexes like NURF (nucleosome remodelling factor) require histone tails to function properly⁶⁶.

Additionally, could the interdependency of Swi/Snf remodelling complexes with HATs be due, in part, to the fact that each is capable of leaving different covalent marks on the chromatin fibre as part of its remodelling function? In so doing, are signals laid down on the histone tails that recruit the next remodelling complex? It is known, for example, that Gcn5-containing HAT complexes are recruited to specific promoters in yeast after the recruitment of Swi/Snf^{48,49}. Could this recruitment partly be caused by the ability of Swi/Snf to leave a mark for HATs to see? It will be of interest to determine whether any nucleosome remodelling complexes contain enzymatic activities that leave covalent marks on histones.

Parallels in nature and conclusions

Like chromatin, microtubules are polymers composed of highly conserved subunits, (α - and β -tubulin) that heterodimerize to form the repeating unit of this cytoskeletal fibre. Tubulins also have 'tail' domains that are decorated by a diverse array of post-translational modifications, some of which are in common with histones (acetylation and phosphorylation)^{67,68}. Like histone tails, the microtubule tails, which are located in the C terminus, also lack a defined structure at atomic resolution⁶⁹ but are known to be recognized by microtubule-associated proteins (MAPs)—polypeptides thought to impart dynamic features to the polymer. We wonder whether unique combinations of post-translational modifications exist on tubulins that modulate their function as in chromatin. The apparent parallels between these two types of cellular polymers are intriguing and suggest that a general theme is used by nature to regulate the dynamics of large polymers. In both cases, the complexity and potential redundant nature of these covalent marks may underlie the general difficulty in obtaining clear phenotypes in mutational analyses of known modification sites.

In summary, the large network of post-translational modifications that decorates histone tails appears to represent a mechanism for differential regulation of chromatin activity in several distinct biological settings. The histone code described here is by no means deciphered, and we have begun to consider the staggering possi-

bility that every amino acid in histone tails has specific meaning and is part of the vocabulary of the overall code. The realization that histone variants, such as the centromere-associated protein CENP-A, exist in special chromosomal locations adds yet another level of variation to the chromatin fibre and the histone code. The recent track record suggests the continued need to: (1) identify and map the sites for the complete dictionary of covalent histone modifications in all histones; (2) mutate and identify phenotypes associated with each of the modifications, singly and in combination; (3) identify and characterize the enzymes systems that add or subtract these modifications; (4) determine how complexes containing these activities are recruited to key genomic targets; and (5) learn how these covalent marks specify interactions with downstream partners or modulate higher-order structures. After a long incubation period, interest in covalent modifications of histones is at an all-time high. Understanding the rules and the consequences of this histone code is likely to impact on many, if not all, DNA-templated process with far-reaching implications for human biology and disease. □

1. Luger, K. & Richmond, T. J. The histone tails of the nucleosome. *Curr. Opin. Genet. Dev.* **8**, 140–146 (1998).
2. Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryotic chromosome. *Cell* **98**, 285–294 (1999).
3. van Holde, K. E. in *Chromatin* (ed. Rich, A.) 111–148 (Springer, New York, 1988).
4. Wolffe, A. P. & Hayes, J. Chromatin disruption and modification. *Nucleic Acids Res.* **27**, 711–720 (1999).
5. Hecht, A., Laroche, T., Strahl-Bolsinger, S., Gasser, S. M. & Grunstein, M. Histone H3 and H4 N-termini interact with SIR3 and SIR4 proteins: a molecular model for the formation of heterochromatin in yeast. *Cell* **80**, 583–592 (1995).
6. Edmondson, D. G., Smith, M. M. & Roth, S. Y. Repression domain of the yeast global repressor Tup1 interacts directly with histones H3 and H4. *Genes Dev.* **10**, 1247–1259 (1996).
7. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
8. Hansen, J. C., Tse, C. & Wolffe, A. P. Structure and function of the core histone N-termini: more than meets the eye. *Biochemistry* **37**, 17637–17641 (1998).
9. Mizzen, C. et al. Signaling to chromatin through histone modifications: how clear is the signal? *Cold Spring Harb. Symp. Quant. Biol.* **63**, 469–481 (1998).
10. Turner, B. M. Decoding the nucleosome. *Cell* **75**, 5–8 (1993).
11. Lopez-Rodas, G. et al. Histone deacetylase. A key enzyme for the binding of regulatory proteins to chromatin. *FEBS Lett.* **317**, 175–180 (1993).
12. Loidl, P. Histone acetylation: facts and questions. *Chromosoma* **103**, 441–449 (1994).
13. Toderia, V., Sendra, R. & Perez-Ortin, J. E. The role of histones and their modifications in the informative content of chromatin. *Experientia* **49**, 780–788 (1993).
14. Grunstein, M. Histone acetylation in chromatin structure and transcription. *Nature* **389**, 349–352 (1997).
15. Struhl, K. Histone acetylation and transcriptional regulatory mechanisms. *Genes Dev.* **12**, 599–606 (1998).
16. Thorne, A. W., Kmicki, D., Mitchelson, K., Sautiere, P. & Crane-Robinson, C. Patterns of histone acetylation. *Eur. J. Biochem.* **193**, 701–713 (1990).
17. Kuo, M. H. et al. Transcription-linked acetylation by Gcn5p of histones H3 and H4 at specific lysines. *Nature* **383**, 269–272 (1996).
18. Grant, P. A. et al. Expanded lysine acetylation specificity of Gcn5 in native complexes. *J. Biol. Chem.* **274**, 5895–5900 (1999).
19. Zhang, W., Bone, J. R., Edmondson, D. G., Turner, B. M. & Roth, S. Y. Essential and redundant functions of histone acetylation revealed by mutation of target lysines and loss of the Gcn5p acetyltransferase. *EMBO J.* **17**, 3155–3167 (1998).
20. Rojas, J. R. et al. Structure of Tetrahymena GCN5 bound to coenzyme A and a histone H3 peptide. *Nature* **401**, 93–98 (1999).
21. Tanner, K. G. et al. Catalytic mechanism and function of invariant glutamic acid 173 from the histone acetyltransferase GCN5 transcriptional coactivator. *J. Biol. Chem.* **274**, 18157–18160 (1999).
22. Trievel, R. C. et al. Crystal structure and mechanism of histone acetylation of the yeast GCN5 transcriptional coactivator. *Proc. Natl Acad. Sci. USA* **96**, 8931–8936 (1999).
23. Clements, A. et al. Crystal structure of the histone acetyltransferase domain of the human PCAF transcriptional regulator bound to coenzyme A. *EMBO J.* **18**, 3521–3532 (1999).
24. Lin, Y., Fletcher, C. M., Zhou, J., Allis, C. D. & Wagner, G. Solution structure of the catalytic domain of GCN5 histone acetyltransferase bound to coenzyme A. *Nature* **400**, 86–89 (1999).
25. Sternglanz, R. & Schindelin, H. Structure and mechanism of action of the histone acetyltransferase gcn5 and similarity to other N-acetyltransferases. *Proc. Natl Acad. Sci. USA* **96**, 8807–8808 (1999).
26. Kimura, A. & Horikoshi, M. How do histone acetyltransferases select lysine residues in core histones? *FEBS Lett.* **431**, 131–133 (1998).
27. Turner, B. M. & O'Neill, L. P. Histone acetylation in chromatin and chromosomes. *Semin. Cell Biol.* **6**, 229–236 (1995).
28. Annunziato, A. T. in *The Nucleus* (ed. Wolffe, A. P.) 31–56 (JAI, Greenwich, Connecticut, 1995).
29. Allis, C. D., Chicoine, L. G., Richman, R. & Schulman, I. G. Deposition-related histone acetylation in micronuclei of conjugating Tetrahymena. *Proc. Natl Acad. Sci. USA* **82**, 8048–8052 (1985).
30. Sobel, R. E., Cook, R. G., Perry, C. A., Annunziato, A. T. & Allis, C. D. Conservation of deposition-related acetylation sites in newly synthesized histones H3 and H4. *Proc. Natl Acad. Sci. USA* **92**, 1237–1241 (1995).
31. Tyler, J. K. et al. The RCAF complex mediates chromatin assembly during DNA replication and repair. *Nature* **402**, 555–560 (1999).
32. Bradbury, E. M. Reversible histone modifications and the chromosome cell cycle. *Bioessays* **14**, 9–16 (1992).
33. Koshland, D. & Strunnikov, A. Mitotic chromosome condensation. *Annu. Rev. Cell Dev. Biol.* **12**, 305–333 (1996).
34. Mahadevan, L. C., Willis, A. C. & Barratt, M. J. Rapid histone H3 phosphorylation in response to growth factors, phorbol esters, okadaic acid, and protein synthesis inhibitors. *Cell* **65**, 775–783 (1991).
35. Thomson, S., Mahadevan, L. C. & Clayton, A. L. MAP kinase-mediated signalling to nucleosomes and immediate-early gene induction. *Semin. Cell Dev. Biol.* **10**, 205–214 (1999).
36. Chadee, D. N. et al. Increased Ser-10 phosphorylation of histone H3 in mitogen-stimulated and oncogene-transformed mouse fibroblasts. *J. Biol. Chem.* **274**, 24914–24920 (1999).
37. Sassone-Corsi, P. et al. Requirement of Rsk-2 for epidermal growth factor-activated phosphorylation of histone H3. *Science* **285**, 886–891 (1999).
38. De Cesare, D., Jacquot, S., Hanauer, A. & Sassone-Corsi, P. Rsk-2 activity is necessary for epidermal growth factor-induced phosphorylation of CREB protein and transcription of c-fos gene. *Proc. Natl Acad. Sci. USA* **95**, 12202–12207 (1998).
39. Thomson, S. et al. The nucleosomal response associated with immediate-early gene induction is mediated via alternative MAP kinase cascades: MSK1 as a potential histone H3/HMG-14 kinase. *EMBO J.* **18**, 4779–4793 (1999).
40. Jin, Y. et al. JIL-1: a novel chromosomal tandem kinase implicated in transcriptional regulation in *Drosophila*. *Mol. Cell* **4**, 129–135 (1999).
41. Lucchesi, J. C. Dosage compensation in flies and worms: the ups and downs of X-chromosome regulation. *Curr. Opin. Genet. Dev.* **8**, 179–184 (1998).
42. Turner, B. M., Birley, A. J. & Lavender, J. Histone H4 isoforms acetylated at specific lysine residues define individual chromosomes and chromatin domains in *Drosophila* polytene nuclei. *Cell* **69**, 375–384 (1992).
43. von Holt, C. et al. Isolation and characterization of histones. *Methods Enzymol.* **170**, 431–523 (1989).
44. Strahl, B. D., Ohba, R., Cook, R. G. & Allis, C. D. Methylation of histone H3 at lysine 4 is highly conserved and correlates with transcriptionally active nuclei in *Tetrahymena*. *Proc. Natl Acad. Sci. USA* **96**, 14967–14972 (1999).
45. Chen, D. et al. Regulation of transcription by a protein methyltransferase. *Science* **284**, 2174–2177 (1999).
46. Nakajima, T. et al. The signal-dependent coactivator CBP is a nuclear target for pp90RSK. *Cell* **86**, 465–474 (1996).
47. Berger, S. L. Gene activation by histone and factor acetyltransferases. *Curr. Opin. Cell Biol.* **11**, 336–341 (1999).
48. Cosma, M. P., Tanaka, T. & Nasmyth, K. Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. *Cell* **97**, 299–311 (1999).
49. Krebs, J. E., Kuo, M. H., Allis, C. D. & Peterson, C. L. Cell cycle-regulated histone acetylation required for expression of the yeast HO gene. *Genes Dev.* **13**, 1412–1421 (1999).
50. Clark, D. et al. Chromatin structure of transcriptionally active genes. *Cold Spring Harb. Symp. Quant. Biol.* **58**, 1–6 (1993).
51. Roth, S. Y. & Allis, C. D. Chromatin condensation: does histone H1 dephosphorylation play a role? *Trends Biochem. Sci.* **17**, 93–98 (1992).
52. Barratt, M. J., Hazzalin, C. A., Cano, E. & Mahadevan, L. C. Mitogen-stimulated phosphorylation of histone H3 is targeted to a small hyperacetylation-sensitive fraction. *Proc. Natl Acad. Sci. USA* **91**, 4781–4785 (1994).
53. Hendzel, M. J. Mitosis-specific phosphorylation of histone H3 initiates primarily within pericentromeric heterochromatin during G2 and spreads in an ordered fashion coincident with mitotic chromosome condensation. *Chromosoma* **106**, 348–360 (1997).
54. Wei, Y., Yu, L., Bowen, J., Gorovsky, M. A. & Allis, C. D. Phosphorylation of histone H3 is required for proper chromosome condensation and segregation. *Cell* **97**, 99–109 (1999).
55. Goto, H. et al. Identification of a novel phosphorylation site on histone H3 coupled with mitotic chromosome condensation. *J. Biol. Chem.* **274**, 25543–25549 (1999).
56. Sullivan, K. F., Hechenberger, M. & Masri, K. Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere. *J. Cell Biol.* **127**, 581–592 (1994).
57. Hirano, T. SMC-mediated chromosome mechanics: a conserved scheme from bacteria to vertebrates? *Genes Dev.* **13**, 11–19 (1999).
58. De Rubertis, F. et al. The histone deacetylase RPD3 counteracts genomic silencing in *Drosophila* and yeast. *Nature* **384**, 589–591 (1996).
59. Braunstein, M., Sobel, R. E., Allis, C. D., Turner, B. M. & Broach, J. R. Efficient transcriptional silencing in *Saccharomyces cerevisiae* requires a heterochromatin histone acetylation pattern. *Mol. Cell Biol.* **16**, 4349–4356 (1996).
60. Dhalluin, C. et al. Structure and ligand of a histone acetyltransferase bromodomain. *Nature* **399**, 491–496 (1999).
61. Winston, F. & Allis, C. D. The bromodomain: a chromatin-targeting module? *Nature Struct. Biol.* **6**, 601–604 (1999).
62. Pawson, T. Protein modules and signalling networks. *Nature* **373**, 573–580 (1995).
63. Roberts, S. M. & Winston, F. Essential functional interactions of SAGA, a *Saccharomyces cerevisiae* complex of Spt, Ada, and Gcn5 proteins, with the Snf/Swi and Srb/mediator complexes. *Genetics* **147**, 451–465 (1997).
64. Biggar, S. R. & Crabtree, G. R. Continuous and widespread roles for the Swi–Snf complex in transcription. *EMBO J.* **18**, 2254–2264 (1999).
65. Sudarsanam, P., Cao, Y., Wu, L., Laurent, B. C. & Winston, F. The nucleosome remodeling complex, Snf/Swi, is required for the maintenance of transcription in vivo and is partially redundant with the histone acetyltransferase, Gcn5. *EMBO J.* **18**, 3101–3106 (1999).
66. Georgel, P. T., Tsukiyama, T. & Wu, C. Role of histone tails in nucleosome remodeling by *Drosophila* NURF. *EMBO J.* **16**, 4717–4726 (1997).
67. Luduena, R. F. Multiple forms of tubulin: different gene products and covalent modifications. *Int. Rev. Cytol.* **178**, 207–275 (1998).
68. Luduena, R. F., Banerjee, A. & Khan, I. A. Tubulin structure and biochemistry. *Curr. Opin. Cell Biol.* **4**, 53–57 (1992).
69. Nogales, E., Whittaker, M., Milligan, R. A. & Downing, K. H. High-resolution model of the microtubule. *Cell* **96**, 79–88 (1999).

Acknowledgements

Research surrounding this topic in our laboratory is supported by grants from the NIH to C.D.A. and B.D.S. We wish to thank current laboratory members, especially C. A. Mizzen, for critical review of this manuscript, and T. K. Archer and C. L. Smith for kindly sharing unpublished data.

Correspondence and requests for materials should be addressed to C.D.A. (e-mail: allis@virginia.edu).

Crystal structure of the hereditary haemochromatosis protein HFE complexed with transferrin receptor

Melanie J. Bennett*, José A. Lebrón* & Pamela J. Bjorkman†

* Division of Biology 156-29, † Howard Hughes Medical Institute, California Institute of Technology, Pasadena, California 91125, USA

HFE is related to major histocompatibility complex (MHC) class I proteins and is mutated in the iron-overload disease hereditary haemochromatosis. HFE binds to the transferrin receptor (TfR), a receptor by which cells acquire iron-loaded transferrin. The 2.8 Å crystal structure of a complex between the extracellular portions of HFE and TfR shows two HFE molecules which grasp each side of a twofold symmetric TfR dimer. On a cell membrane containing both proteins, HFE would 'lie down' parallel to the membrane, such that the HFE helices that delineate the counterpart of the MHC peptide-binding groove make extensive contacts with helices in the TfR dimerization domain. The structures of TfR alone and complexed with HFE differ in their domain arrangement and dimer interfaces, providing a mechanism for communicating binding events between TfR chains. The HFE–TfR complex suggests a binding site for transferrin on TfR and sheds light upon the function of HFE in regulating iron homeostasis.

Hereditary haemochromatosis (HH) is an autosomal recessive disorder that is more common than cystic fibrosis, muscular dystrophy and phenylketonuria combined (reviewed in refs 1, 2). Roughly 1 in 300 Caucasians have HH, which gives an estimated carrier frequency of about 1 in 9. HH patients chronically absorb a small excess of iron, resulting in body iron stores of up to ten times the normal levels. Excess iron is deposited in virtually every major organ, particularly in the parenchymal cells of the liver, pancreas and heart, leading to multi-organ failure.

The gene mutated in HH encodes a 348-residue type I transmembrane glycoprotein, HFE, which is homologous to class I MHC molecules and associates with the class I light chain β 2-microglobulin (β 2m)³. Unlike classical class I MHC molecules, which function in the immune system by presenting peptide antigens to T cells⁴, HFE does not bind peptides or perform any known immune function⁵. Most HH patients are homozygous for a mutation (845G to A) that converts Cys 260 (numbering begins at the first residue of the mature protein⁵) to a tyrosine, eliminating a disulphide bond in the α 3 domain of HFE and preventing association with β 2m and cell-surface expression in cell-culture models^{6,7}. A second mutation, which converts HFE His 41 to aspartate (His41Asp), is found on the other chromosome in about 70% of HH patients who are heterozygous for the Cys260Tyr mutation³. This substitution does not prevent β 2m association or cell-surface expression^{6,7}. The 2.6 Å resolution crystal structure of the soluble HFE ectodomain verified that HFE resembles class I MHC molecules⁵. In both types of protein, the α 1– α 2 superdomain (residues 1–180) forms a platform comprising an eight-stranded antiparallel β -sheet topped by two α -helices, which is positioned on top of the immunoglobulin constant-like domains α 3 (residues 181–275) and β 2m (Fig. 1a). HFE has a narrower and shallower version of the class I peptide-binding groove. Figure 1a highlights the residues affected by the HH mutations and an unusual cluster of four histidine residues resembling a metal-binding site.

Consistent with its role in an iron-overload disorder, HFE binds to the transferrin receptor (TfR)^{8,9}, a macromolecule central to iron homeostasis^{10,11}. TfR is a homodimeric type II transmembrane glycoprotein that binds ferric-iron-loaded transferrin (Fe-Tf) at the slightly basic pH of the cell surface (~pH 7.4). Fe-Tf–TfR complexes are endocytosed into acidic vesicles, where iron is released from Tf. Iron-free Tf (apo-Tf) remains bound to TfR at

acidic pH, and is recycled to the cell surface where the basic pH of the bloodstream triggers its dissociation^{10,11}. In solution, HFE binds TfR tightly at the pH of the cell surface, but not at pH 6, suggesting that HFE dissociates from TfR in acidified endosomes^{5,12}. Biochemical studies showed that HFE and Fe-Tf can bind simultaneously to TfR to form a ternary complex⁵ consisting of one Fe-Tf and one HFE bound to a TfR homodimer¹³, and that HFE inhibits the TfR–Tf interaction^{8,14} by binding at or near the Tf-binding site on TfR¹³. The 3.2 Å crystal structure of the homodimeric TfR ectodomain¹⁵ revealed three domains: a domain resembling amino- and carboxypeptidases (protease-like domain; residues 121–188 and 384–606), an apical domain (residues 189–383) and a helical domain involved in dimerization consisting of six helices arranged as a four-helix bundle (residues 607–760) (Fig. 1b). In each polypeptide chain of the TfR dimer, the three domains form a cleft that was suggested to bind one molecule of Tf⁵.

Here we describe the 2.8 Å crystal structure of a complex between the ectodomains of human HFE and TfR. The complex has twofold symmetry with a 2:2 stoichiometry, such that one HFE contacts each polypeptide chain of the TfR homodimer. The relative orientations of both molecules indicate that HFE and TfR associate on the same membrane, rather than between opposing membranes. Conformational changes in TfR induced by HFE binding are discussed in light of the role of HFE in maintenance of iron homeostasis.

Overview of the structure

The HFE–TfR structure was determined by multiple isomorphous replacement with anomalous scattering (MIRAS) aided by threefold noncrystallographic symmetry (NCS) averaging (see Methods). The model derived after refinement to 2.8 Å (Table 1) is shown in Fig. 1c. The complex consists of one HFE molecule bound to each chain of the TfR homodimer to form a twofold symmetric complex. There are extensive contacts between the two polypeptide chains in the TfR dimer, but no contacts between the two HFE molecules. The primary intermolecular contacts involve the HFE α 1– α 2 platform and the helical domain of TfR. The α 3 and β 2m domains of HFE do not contact TfR. Temperature factors for the HFE–TfR model increase with the distance from the centre of the complex (see Methods), indicating that TfR and the region of HFE that contacts TfR form a relatively stable core, whereas the α 3 domain and β 2m are more mobile.

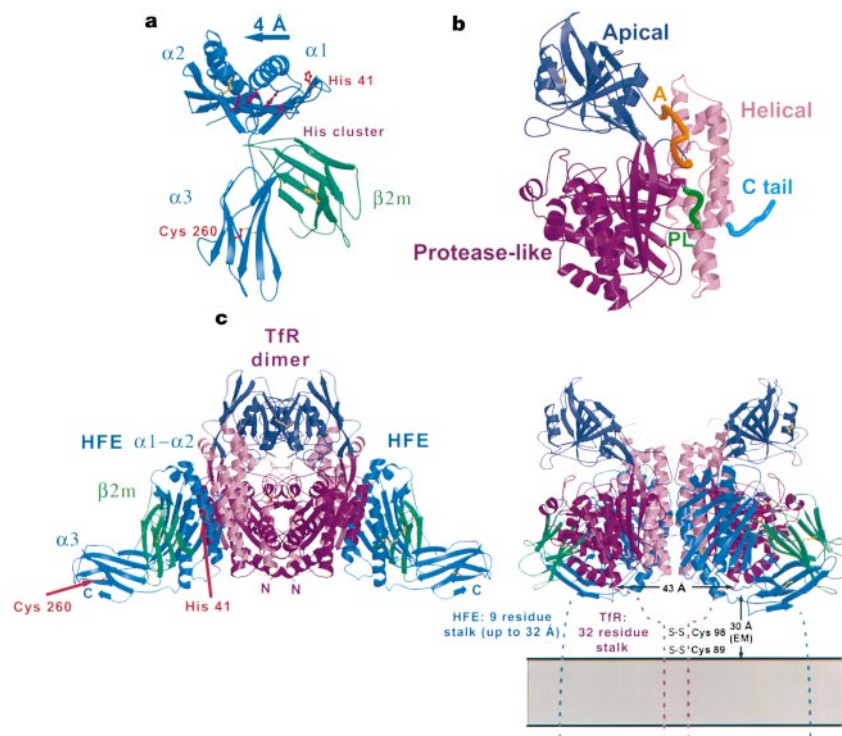


Figure 1 Ribbon diagrams of HFE, TfR and HFE-TfR structures. **a**, HFE⁵ (PDB code 1A6Z). Residues substituted in HH mutations (Cys 260 and His 41)⁵ and a cluster of histidines (residues 87, 89, 94 and 123) are highlighted. An arrow indicates the inward displacement of $\alpha 1$ domain helix as compared with the analogous class I MHC helix. **b**, TfR monomer from homodimer structure¹⁵ (made using coordinates provided by C. M.

Lawrence and S. C. Harrison). A, apical loop (residues 312–328); PL, protease-like loop (residues 469–476); C tail, C-terminal tail (residues 750–760). **c**, Two views of the HFE-TfR structure related by a 90° rotation about the vertical axis. Chain termini nearest the predicted transmembrane region (C terminus for HFE heavy chain; N terminus for TfR) are labelled (left). The membrane bilayer is represented by a grey box (right).

The overall orientations of TfR and HFE, as well as the locations of the ectodomain termini on the same face of the complex, indicate that the molecules interact while anchored to the plane of a common membrane, eliminating the possibility that HFE on one membrane binds to TfR on another membrane. As shown in Fig. 1c, the ectodomain termini closest to the membrane (amino termini for TfR chains, carboxy termini for HFEs) are located at roughly the same level. Taking into account the ~30 Å length of the TfR stalk derived from electron microscopy (EM) studies¹⁶, and assuming that the ~9 residues remaining in the HFE extracellular region could span up to 32 Å in an extended conformation¹⁷, we predict that the portion of the HFE-TfR complex that was crystallized is situated ~30 Å from the plane of a common membrane (Fig. 1c). The HFE molecules are oriented with their long axes roughly parallel to the predicted plane of the membrane in a ‘lying down’ orientation, as compared with the ‘standing up’ orientation assumed to be relevant for MHC molecules, which extend their peptide-binding grooves away from the cell surface to bind T-cell receptors (TCRs) on other cells⁴.

HFE-TfR interface

HFE and TfR interact over a large interface, such that a total of 2,000 Å² (1,000 Å² per subunit; calculated as described¹⁸) in solvent-accessible surface area is buried when one HFE molecule binds to a TfR monomer. The surface area buried is larger than that buried in several other macromolecular complexes in which, like HFE and TfR, the individual components are stable as monomers (mean areas buried per subunit: 780–850 Å²)¹⁸.

The primary interaction between HFE and TfR involves the $\alpha 1$ and $\alpha 2$ domain helices of HFE and two helices within the helical domain of TfR (Fig. 2a, b; Table 2). At the core of the interface is a three-helix bundle consisting of the $\alpha 1$ helix from HFE and helical domain helices 1 and 3 from TfR (Fig. 2c). In crystals of uncom-

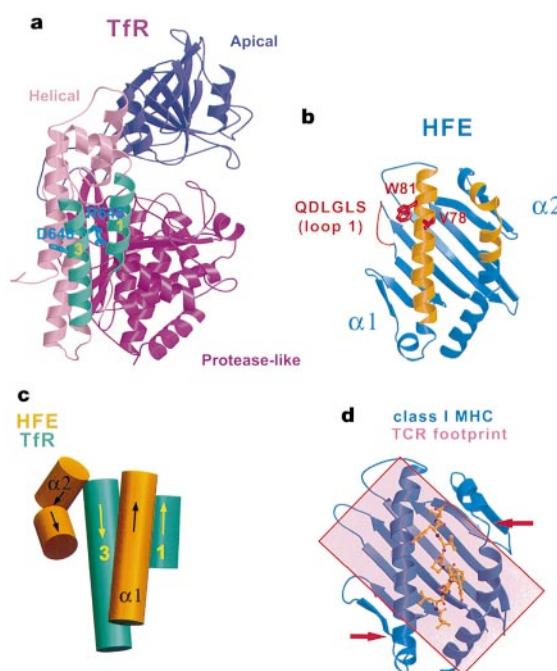


Figure 2 The HFE-TfR interface. Cut away views of a TfR monomer (**a**) and the HFE platform (**b**) from the HFE-TfR structure are shown with regions containing residues that form contacts highlighted in gold (HFE) and aqua (TfR). Side chains of residues identified by site-directed mutagenesis to affect ligand binding are highlighted on TfR in cyan (Fe-Tf binding²⁹) and on HFE in red (TfR binding¹²). **c**, Representation of HFE-TfR interface showing the three-helix bundle and kinked portion of the HFE $\alpha 2$ helix. **d**, TCR footprint (pink box) on the structure of a class I MHC molecule. Arrows indicate the ‘high points’ on the $\alpha 1$ and $\alpha 2$ helices (Based on Fig. 5 in ref. 4).

Table 1 Data collection, phasing and refinement statistics for HFE–TfR

Data Set	Resolution (Å)	Complete (%) [*]	$R_{\text{merge}}^{\dagger}$ (%)	I/σ	R.m.s. $f_o/E^{\ddagger}$
Native	2.8 (2.9–2.8)	98.1 (99.8)	7.0 (34.1)	18.0 (3.6)	
K ₂ PtCl ₄	3.5 (3.6–3.5)	93.3 (78.6)	9.1 (38.2)	10.9 (2.6)	1.8
PIPS	4.0 (4.1–4.0)	85.9 (85.6)	14.0 (35.8)	5.8 (2.2)	1.6
Uranyl acetate	4.2 (4.4–4.2)	95.3 (81.0)	10.3 (26.9)	10.7 (3.9)	0.8
Refinement					
Resolution (Å)	30–2.8		R.m.s. $\Delta\phi$ all NCS residues (deg) [¶]		0.4, 1.0
Reflections (working set) $ F > 0$	106,875		R.m.s. $\Delta\psi$ all NCS residues (deg) [¶]		0.4, 0.7
Reflections (test set) $ F > 0$	11,868				
R_{free} (%)	26.4		Number of nonhydrogen atoms		
R_{cryst} (%)	22.8		Protein		24,258
			Water		9
			Ca ²⁺		3
			Carbohydrate		42
			Glycerol		6
R.m.s.d. from ideality			Ramachandran plot (non-glycine)		
Bond lengths (Å)	0.007		Most favoured		84.1
Bond angles (deg)	1.34		Additional allowed		14.2
			Generously allowed		1.0
			Disallowed		0.7

Values in parentheses indicate the high-resolution shells.

^{*} Complete indicates number of independent reflections/total theoretical number.

[†] $R_{\text{merge}}(I) = (\sum_i |I_i - \langle I \rangle|) / \sum_i I_i$, where I_i is the i th observation of the intensity of the hkl reflection and $\langle I \rangle$ is the mean intensity from multiple measurements of the h, k, l reflection.

[‡] R.m.s. f_o/E (phasing power), where f_o is the heavy-atom structure-factor amplitude and E is the residual lack of closure error.

[§] PIP, di-*m*-iodobis(ethylenediamine)platinum nitrate. The overall figure of merit (acentric reflections) including the anomalous signal is 0.44.

^{||} $R_{\text{cryst}}(F) = \sum_i |F_{\text{obs}}(h) - F_{\text{calc}}(h)| / \sum_i |F_{\text{obs}}(h)|$, where $|F_{\text{obs}}(h)|$ and $|F_{\text{calc}}(h)|$ are the observed and calculated structure-factor amplitudes for the h, k, l reflection. R_{free} is calculated over reflections in a test set not included in atomic refinement.

[¶] Statistics for NCS-related residues refer to differences relative to chains G–I for the two other half complexes (chains A–C and D–F) in the crystallographic asymmetric unit.

plexed HFE⁵, the $\alpha 1$ helix was involved in a crystal contact in which the $\alpha 1$ and $\alpha 2$ helices of a symmetry-related HFE molecule mimicked some features of the HFE–TfR interface. In the HFE–TfR co-crystals, the HFE $\alpha 1$ helix and TfR helix 3 are antiparallel to each other and interact over nearly the entire lengths of both helices (~5 turns). The antiparallel helices lack heptad repeats with hydrophobic residues pointing into the interface, a feature of α -helical coiled coils¹⁹, and do not form a coiled coil similar to the antiparallel coil in seryl-tRNA synthetase²⁰. Helix 1 from TfR, which is shorter than the other two helices in the bundle, interacts with the C-terminal region of the HFE $\alpha 1$ helix in a parallel arrangement (Fig. 2c). This short helix pair has a hydrophobic core consisting of Leu 619 and Val 622 from TfR packed against Val 78 and Trp 81 from HFE (Fig. 3a). The importance of these interactions for HFE binding to TfR is underscored by the finding that site-directed mutants of HFE at positions 78 and 81 either show no detectable binding to TfR (Val78Ala) or bind with an affinity reduced by ~5,000-fold (Trp81Ala)¹². These residues are also near the positions affected by two mutations implicated in HH, Ile83Thr and Gly71Arg (ref. 21).

Additional contacts between HFE and TfR involve Leu 22 in a flexible $\alpha 1$ domain loop (loop 1) of HFE (Fig. 2b) interacting with helix 1 of the TfR helical domain, and residues near the kink in the $\alpha 2$ helix of HFE (residues 146–156) (Fig. 3a), which interact with helix 3 of the TfR helical domain (Fig. 2c). The involvement of Leu 22 in loop 1 in the binding interface is consistent with mutagenesis results, in which substitution of the loop containing this residue resulted in a ~10-fold reduction in affinity for TfR¹². Residues near the kink in the $\alpha 2$ helix of HFE have not been mutated, but the amount of solvent-accessible surface area buried by these residues (Table 2) indicates that this region may contribute considerably to binding. In addition, Arg 629 in helix 2 of the TfR helical domain makes several polar interactions with residues in the $\alpha 1$ and $\alpha 2$ helices of HFE (Fig. 3a).

The use of the $\alpha 1$ and $\alpha 2$ domain helices of HFE as a recognition surface is reminiscent of the interaction between MHC molecules and TCRs, but the details of the HFE–TfR interaction differ substantially from class I MHC interactions with TCRs. Whereas both TCRs and TfR bind across the $\alpha 1$ – $\alpha 2$ domain helices of their class I or class-I-like ligands, class-I-restricted TCRs avoid the ‘high points’ of the $\alpha 1$ – $\alpha 2$ helices to bind with a diagonal interaction

(Fig. 2d; reviewed in ref. 4), whereas the ‘high point’ or kink in the $\alpha 2$ helix of HFE is a major contact point for TfR (Fig. 2b; Table 2). In addition, TCRs recognize a peptide bound between the class I $\alpha 1$ – $\alpha 2$ helices, whereas the counterpart of the peptide-binding groove in HFE is narrowed and does not bind a peptide or other ligand⁵. The narrowing of the HFE groove is the result of a ~4 Å inward translation of the HFE $\alpha 1$ helix relative to the position of this helix in class I molecules (Fig. 1a), and this is the only part of the HFE backbone structure that deviates significantly from that of class I MHC molecules⁵. The distance between the $\alpha 1$ and $\alpha 2$ helices of HFE is critical for its interaction with TfR; thus, if the HFE groove were open as the result of binding a peptide or other small molecule compound, it would prevent the interaction between HFE and TfR seen in the co-crystals.

Conformational changes

The binding of TfR leads to relatively few changes in the structure of HFE. The major main-chain conformational changes in complexed versus uncomplexed HFE⁵ are in the loops joining $\alpha 1$ domain strands 1 and 2, 3 and 4, the loop after strand 4, and the high point in the $\alpha 2$ helix, all of which are near or at the HFE–TfR interface. Several side chains of HFE at the interface are also reoriented (residues 64, 67, 74, 81, 85 and 152). In comparison, the structures of uncomplexed¹⁵ and HFE-complexed TfR differ substantially. The uncomplexed TfR structure was determined at a pH between 6.5 and 7 (ref. 15), whereas the HFE–TfR complex was determined at pH 8. Although TfR undergoes pH-induced changes as the pH is lowered^{10,11,22,23}, we assume that the changes in complexed TfR result from HFE binding rather than pH differences because studies of TfR-facilitated release of iron from Fe–Tf suggest that TfR changes occur at pH values below 6 (ref. 24).

The first type of conformational change in complexed TfR involves the relative arrangement of domains in the TfR monomer. The backbone structures of the individual domains in TfR are similar in uncomplexed¹⁵ and complexed TfR (the root mean square deviation (r.m.s.d.) values between C α atoms after superposition of individual domains are 0.53 Å (protease-like), 0.52 Å (apical) and 0.85 Å (helical), calculated after eliminating residues in loops that differ more than 2.0 Å). However, the helical domain in uncomplexed TfR is rotated with respect to the remainder of the molecule (Figs 3c, 4a). The movement is supported by an r.m.s.d. of ~2.5 Å

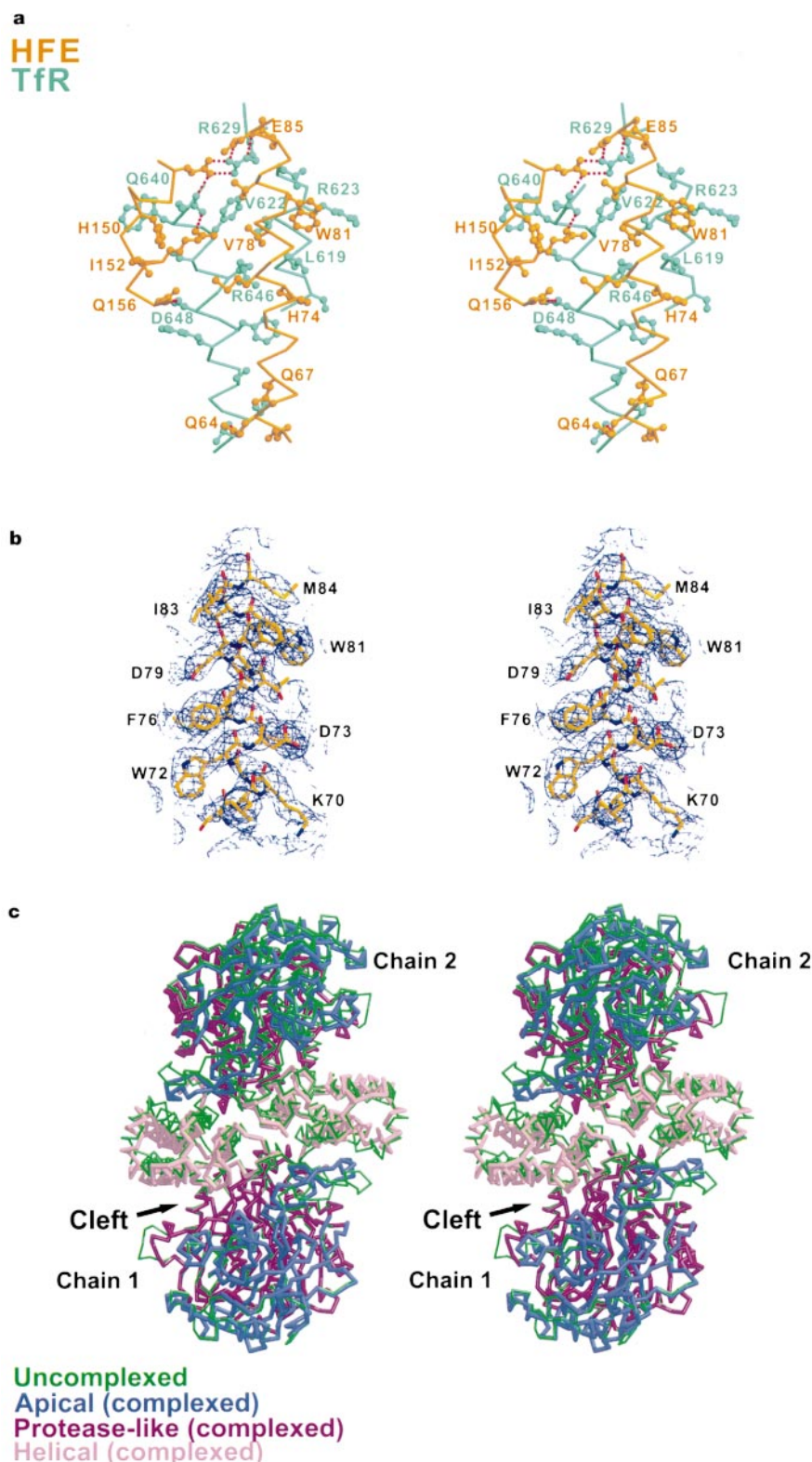


Figure 3 Stereo views of the HFE–TfR structure. **a**, View showing interacting sidechains at the HFE–TfR interface. Intermolecular hydrogen bonds and salt bridges are indicated by dashed lines. Orientation is the same as in Fig. 2c. **b**, View of the 2.8 Å experimental electron-density map contoured at 1.0 σ (calculated with observed amplitudes and NCS-

averaged and phase-extended MIRAS phases) superimposed on the final model in the region of the HFE α 1 helix. **c**, Overlay (top view) of uncomplexed¹⁵ and complexed TfR structures after superposition of C α atoms in the protease-like and apical domains of chain 1.

calculated for the C α atoms of the helical domain after superposing the apical and protease-like domains of uncomplexed¹⁵ and complexed TfR. The helical domain movement occurs by pivoting about a portion of the loop (residues 698–710) following helix 4, which itself moves very little (Fig. 4a). The rigid body differences in helical domain C α positions range from essentially none in the pivot point to a maximum of 4.6 Å in helix 3, where HFE binds.

As a result of the movement of the helical domain, an inter-domain cleft formed by elements of the apical and protease-like domains and helices 1 and 5 of the helical domain in uncomplexed TfR¹⁵ appears relatively closed in complexed TfR (Figs 3c, 4b). The altered shape of the cleft is due in part to rearrangement of a long loop between helices 7 and 8 of the protease-like domain (residues 519–532). This loop may be important for Tf binding, as it contains Trp 528, which is solvent exposed in uncomplexed TfR¹⁵, but buried in TfR complexed with HFE. The HFE loop containing His 41 is near this TfR loop, but makes no direct contacts, consistent with biochemical studies showing that alteration of His 41 and nearby residues does not affect affinity for TfR¹².

Other conformational changes in complexed TfR involve the dimer interface. Both uncomplexed¹⁵ and complexed TfR have extensive dimer interfaces that bury 2,550 Å² and 2,960 Å² of surface area per subunit, respectively. Both dimers are twofold symmetric, but there is an ~1° reorientation of the twofold axis that relates the TfR polypeptide chains. As a result of the reorientation, the two TfR chains in each dimer have a different relative arrangement (after superposition of the apical and protease-like domains of one TfR chain in the uncomplexed¹⁵ and complexed structures, the same domains on the other chain have an r.m.s.d. of 1.6 Å between C α atoms) (Fig. 3c). The dimer interface of both TfR structures is

formed mainly by contacts involving the helical domain and three loops: the C-terminal tail ('C tail'; residues 750–760); a loop in the apical domain ('A loop'; residues 312–328); and a loop in the protease-like domain ('PL loop'; residues 469–476) (Fig. 4c). The dimer interfaces differ because of three types of movements: movement of the helical domain as a rigid body (Fig. 4a); local rearrangement of the main chain in the A, PL and C-tail loops (Fig. 4d); and re-orientation of side chains. We assume that the side-chain and loop conformational changes at the TfR dimer interface are triggered by rigid-body rotation of the helical domain upon binding HFE.

Dependence of HFE–TfR binding on pH

We previously demonstrated that HFE binds TfR with nanomolar affinity at pH 7.5, but weakly or not at all at pH 6.0 (refs 5, 12). Histidines are likely candidates to mediate pH-dependent interactions near neutrality because their pK_a is ~6.5. On HFE, neither His 41 nor the residues in the histidine cluster (Fig. 1a) are at the interface with TfR (Table 2), confirming that they do not mediate pH-dependent binding¹². Thus, other residues on HFE and/or TfR must be responsible for the sharp pH dependence. Two possibilities on HFE are His 74 and His 150, both of which are involved in hydrophobic contacts with TfR residues (Table 2) that might be disrupted if they were protonated at acidic pH. His 150 is conserved in HFE sequences from other species²⁵.

TfR participates in other pH-dependent interactions, in which it facilitates release of iron from Fe–Tf and binds apo–Tf at acidic but not basic pH^{10,11,22}, and it undergoes a conformational change resulting in self-association at pH <6 in the absence of Tf²³. Thus, it is possible and perhaps likely that changes in TfR, rather than HFE, mediate the pH-dependent HFE–TfR interaction. Four histidine residues, His 684 and His 475 from each polypeptide chain, are clustered in the TfR dimer interface at distances of 5–8 Å in uncomplexed TfR¹⁵ (Fig. 4e). In the structure of complexed TfR, His 475 is shifted (Fig. 4d) and forms a hydrogen bond with the partner His 475 across the dimer interface (Fig. 4e). Upon protonation at acidic pH, the His 475 residues could not exist in such close proximity, and their repulsion might produce a rearranged TfR structure that no longer binds HFE. If Fe–Tf binding at pH 7.4 also brings these histidines into close proximity, rearrangements in the TfR dimer interface at acidic pH could modulate TfR-facilitated release of iron from Tf^{22,24,26}.

Transferrin binding to TfR

Binding studies using a human/chicken chimaeric TfR suggest that Tf binds to a region corresponding to the helical domain²⁷. Site-directed mutagenesis localizes the binding site to include residues 646–648 (ref. 28) in helix 3 of the helical domain (Fig. 2a). HFE binds to the helical domain of TfR (Fig. 2a) and contacts residues 646 and 648 (Table 2); thus the Tf-mapping studies are compatible with the suggestion derived from inhibition studies that HFE and Tf bind to the same or overlapping sites on TfR¹³. The interdomain cleft formed by portions of the three TfR domains has also been proposed to be the Tf-binding site¹⁵ (Fig. 4b). As Fe–Tf is a large molecule (~90 Å × 50 Å × 40 Å; measured using the structure of lactoferrin²⁹; protein data bank (PDB) code 1LFG), it could simultaneously contact the interdomain cleft and the HFE-binding site on the TfR helical domain.

Stoichiometry of the TfR–HFE complex

Studies at neutral or slightly basic pH showed that TfR and HFE interact with 2:1 stoichiometry (one TfR homodimer binding to one HFE) at micromolar concentrations in solution⁵ (L. Joss, J.A.L., P.J.B. and D. G. Myszk, unpublished observations) and in transfected cell lysates¹⁴. However, the TfR–HFE complex structure has a 2:2 stoichiometry, showing that the second HFE-binding site on TfR can be occupied at the millimolar concentration of the co-crystals.

Table 2 Interactions between HFE and TfR

HFE			TfR		
Residue	Location	% ASA	Residue	Location	% ASA
Leu 22	loop 1	7	Ser 616 (Asp)	HD helix 1	3
Leu 22	loop 1		Leu 619 (Leu)	HD helix 1	10
Leu 22	loop 1		Arg 623 (Gly)	HD helix 1	7
Leu 63	α1 helix	4	Thr 657 (Arg)	HD helix 3	4
Gln 64*	α1 helix	3	Thr 658* (Gln)	HD helix 3	4
Gln 67	α1 helix	7	Ser 654 (Glu)	HD helix 3	3
Gln 67	α1 helix		Thr 657	HD helix 3	
Gly 71	α1 helix	2	Phe 650 (Ile)	HD helix 3	7
His 74	α1 helix	9	Ser 616	HD helix 1	
His 74	α1 helix		Leu 619	HD helix 1	
His 74	α1 helix		Arg 646 (Arg)	HD helix 3	5
His 74	α1 helix		Phe 650	HD helix 3	
Met 75	α1 helix	2	Arg 646	HD helix 3	
Val 78	α1 helix	8	Leu 619	HD helix 1	
Val 78	α1 helix		Val 622 (Ile)	HD helix 1	2
Val 78	α1 helix		Tyr 643 (Tyr)	HD helix 3	4
Trp 81	α1 helix	10	Leu 619	HD helix 1	
Trp 81	α1 helix		Val 622	HD helix 1	
Trp 81	α1 helix		Arg 623	HD helix 1	
Trp 81	α1 helix		Asn 626 (Asn)	HD helix 1	5
Thr 82	α1 helix	3	Arg 629 (Ser)	HD helix 2	10
Thr 82	α1 helix		Tyr 643	HD helix 3	
Glu 85*	α1 helix	7	Arg 629*	HD helix 2	
Asn 86*	α1 helix	<1	Arg 629*	HD helix 2	
Glu 146*	α2 helix	6	Arg 629*	HD helix 2	
Glu 146*	α2 helix		Gln 640* (Gln)	HD helix 3	10
His 150	α2 helix	9	Gln 640	HD helix 3	
His 150	α2 helix		Trp 641 (Trp)	HD helix 3	5
His 150	α2 helix		Ser 644 (Ser)	HD helix 3	4
Ile 152	α2 helix	3	Trp 641	HD helix 3	
Arg 153*	α2 helix	2	Gln 640*	HD helix 3	
Gln 156*	α2 helix	6	Asp 648* (Asp)	HD helix 3	1
Gln 156*	α2 helix		Arg 651 (Arg)	HD helix 3	5

HD, helical domain of TfR. Interface residues were identified as described¹⁸. % ASA (accessible surface area), percent of the total interface ASA contributed by each residue. Residues contributing less than 2% of interface ASA are excluded unless they are involved in hydrogen bonds or salt bridges (indicated by asterisks). Pairwise interactions were identified by contact analysis in CNS⁴⁰. Residues in parentheses correspond to the sequence of TfR2, a Tf-binding TfR homologue⁴⁶. Some of the TfR residues critical for binding HFE are not conserved in TfR2.

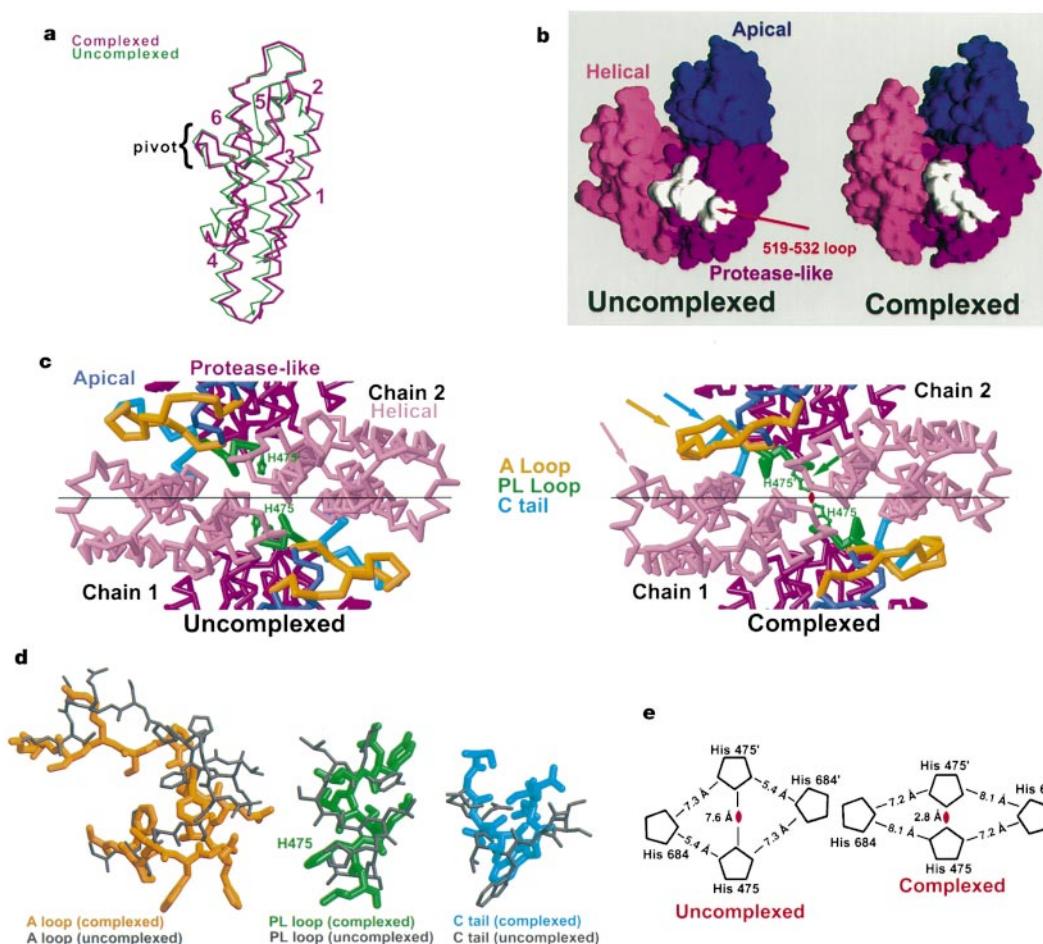


Figure 4 Conformational changes in TfR. **a**, Comparison of the positions of the helical domains in uncomplexed¹⁵ and complexed TfR after superposition of C α atoms in the protease-like and apical domains. Helices are numbered. **b**, GRASP⁴⁷ representations of the solvent-accessible surfaces of TfR monomers in the uncomplexed¹⁵ and complexed structures. **c**, Comparison (top view) of dimer interfaces after superposition of the apical and protease-like domains in chain 1 of the TfR structures. The view for both panels is

Because of the high concentrations in a crystalline environment, a 2:1 TfR–HFE complex is not expected to be observed crystallographically; thus, we cannot ascertain the structural mechanism by which a 2:1 complex is maintained in solution. However, the 2:2 TfR–HFE complex found in the crystals reveals plasticity in the TfR dimer interface that could be involved in reducing the affinity for binding a second HFE. For example, HFE binding to one side of TfR could propagate changes across the TfR dimer interface that make it less favourable for a second HFE to bind. Because the TfR–HFE complex has twofold symmetry, any potential asymmetry in a 2:1 complex must be lost upon binding a second HFE molecule.

The stoichiometry of a TfR–HFE complex on a cell membrane, and how that is influenced by Fe–Tf, remain to be determined. The structural and biochemical evidence that HFE and Fe–Tf bind to the same or overlapping sites on TfR (see above) suggests that TfR with two bound Fe–Tf molecules³⁰ cannot bind HFE. Likewise, TfR–HFE complexes with a 2:2 stoichiometry cannot bind Fe–Tf. As HFE and TfR are tethered to the same membrane, HFE should compete effectively to bind TfR despite the high physiological concentrations of Fe–Tf (micromolar)³¹, thereby forming either 2:2 TfR–HFE complexes or 1:2:1 Tf–TfR–HFE ternary complexes. It is unlikely that all cell-surface TfR is complexed with two HFE molecules, as expression of HFE in transfected cells does not prevent TfR-mediated endocytosis of Fe–Tf even when HFE is overexpressed and Fe–Tf is present at subphysiological concentrations^{14,32}. Thus, a ternary complex of Fe–Tf, TfR and HFE may be involved in Fe–Tf

uptake. Ternary complexes have been observed in solution using purified proteins⁵, in which the stoichiometry was established as 1:2:1 Fe–Tf:TfR:HFE (ref. 13), and in HFE-transfected cell lysates¹⁴.

Conclusions

The binding of HFE to TfR, a key receptor for iron uptake^{10,11}, is undoubtedly central to the function of HFE in regulation of iron homeostasis. HFE binds to TfR in duodenal crypt enterocytes^{33,34}, which regulate the absorption of dietary iron^{1,2}. The association between HFE and TfR has also been seen in placenta⁹ and in cell lines overexpressing HFE^{8,14}, in which HFE associates with TfR throughout the biosynthetic pathway and colocalizes with Tf in intracellular acidic vesicles¹⁴. The HFE–TfR co-crystal structure provides a detailed view of the protein complex that would exist at the pH of the cell surface. Whereas HFE changes relatively little upon complex formation, the TfR structure¹⁵ undergoes rearrangements. The changes, many of which are at the TfR dimer interface, might be extensive enough to propagate across the membrane to the TfR cytoplasmic tail, thereby communicating to cytoplasmic proteins that HFE is bound to TfR. Although HFE is predicted to dissociate from TfR in acidic vesicles^{5,12}, the imprint of HFE-induced structural changes might influence TfR-facilitated release of iron from Fe–Tf at acidic pH (refs 22, 24, 26, 32). Alternatively, HFE may bind TfR only as a means to gain entry into endosomes, where it would then interact with other molecule(s) to regulate iron metabolism. As neither His 41, the site of an HH mutation³, nor a

cluster of histidines on HFE is at the interface with TfR, one or both of these regions of HFE (Fig. 1a) could mediate an interaction with a different molecule at low pH.

The diversity of functions in the family of proteins related to class I MHC molecules³⁵ raises intriguing questions about the ancestral function of the MHC fold. Reminiscent of the HFE–TfR interaction, class I MHC molecules associate with the insulin receptor on the same membrane³⁶. The $\alpha 1$ domain helix of class I molecules comprises at least part of the insulin-receptor binding site, because synthetic peptides corresponding to this region alter glucose metabolism through their effect on insulin receptors³⁷. Likewise, the $\alpha 1$ helix of HFE is the primary binding site for TfR, and a peptide corresponding to this helix can substitute for soluble HFE in cell-based measurements of the HFE-modulated affinity between TfR and Tf (D. M. Penny *et al.*, manuscript in preparation). On the basis of these common recognition properties, the ancestral function of HFE, class I and class-I-related proteins may have involved using the $\alpha 1$ – $\alpha 2$ domain helices to bind to a receptor anchored on the same membrane. □

Methods

Crystallization and data collection

Ectodomains of human TfR (residues 121–760) and human HFE (residues 1–275) complexed with $\beta 2m$; our numbering system for HFE begins with the N terminus of the mature protein and differs from the system that starts with the initial methionine of the signal sequence (for example, ref. 3)) were expressed and purified as described³. Crystals of a 2:2 complex consisting of two HFEs bound to one TfR homodimer (space group C2; $a = 110.4 \text{ \AA}$, $b = 144.4 \text{ \AA}$, $c = 327.1 \text{ \AA}$, $\beta = 93.6^\circ$) were grown in hanging drops containing 7–8% PEG 8000, 200 mM Tris pH 8.0 and 10 mM trimethylamine HCl. The asymmetric unit of the crystals contains one 2:2 complex (the ‘NCS complex’; HFE chains A and D, $\beta 2m$ chains B and E, TfR chains C and F) and one half complex (the ‘crystallographic complex’; HFE chain G, $\beta 2m$ chain H, TfR chain I), which forms a 2:2 complex by interacting with a symmetry mate related by a crystallographic twofold axis. Crystals were transferred stepwise (5% increments to cryoprotectant (artificial mother liquor plus 20% glycerol) and 2.8 Å native data were collected at -170°C at SSRL beamline 9-1. Heavy-atom-derivative data were collected with an RAXIS IV detector and Rigaku rotating anode from cryopreserved crystals soaked in 0.5–1.0 mM heavy-atom solutions for 1–8 days. Data processing was done with the HKL package³⁸.

Structure determination and refinement

Nine platinum sites were identified by interpretation of the K_2PtCl_4 difference Patterson map. Other derivatives were interpreted using difference Fourier. Heavy-atom refinement and calculation of solvent-flattened SIRAS and MIRAS maps (15–3.6 Å) were done using SHARP³⁹. Using manual and automated real-space search methods⁴⁰, we identified the locations of three TfR monomers in skeletonized 3.6 Å SIRAS maps using the coordinates of aminopeptidase (PDB code 1XJO) as a model for the TfR protease-like domain. The remaining TfR domains were located using the coordinates of uncomplexed TfR¹⁵ (provided by C. M. Lawrence and S. C. Harrison), which facilitated identification of density corresponding to three HFE molecules. Threefold NCS subdomain averaging and phase extension to 2.8 Å were carried out with DM⁴¹, which produced unambiguously interpretable maps (Fig. 3b). Refinement was done with CNS⁴² using overall B factor and bulk-solvent corrections and tight NCS restraints (300 kcal mol⁻¹ Å⁻²) for all but TfR residues 204–211. The TfR dimers in the NCS and crystallographic complexes are virtually identical, with an r.m.s.d. value of 0.17 Å for 1,266 superimposed C α atoms. The $\alpha 3$ and $\beta 2m$ domains of one of the HFEs (chains D and E) in the NCS complex are involved in crystal contacts and deviate from the twofold symmetry of the complex. Individual B factors were refined (refinement of grouped B factors resulted in a 1.3% higher R_{free}). Average B values for the domains in the crystallographic complex and the NCS complex, respectively, are: HFE $\alpha 1$ – $\alpha 2$ platform, 66 Å², 76 Å²; HFE $\alpha 3$ domain, 105 Å², 98 Å²; $\beta 2m$, 93 Å², 92 Å²; and TfR, 49 Å², 60 Å². Despite the high B factors, most of the model is well defined in the electron density (Fig. 3b). The model ($R_{\text{free}} = 26.4\%$, $R_{\text{cryst}} = 22.8\%$) consists of 3,018 protein residues in three copies each of HFE (residues 4–275), $\beta 2m$ (residues 1–99) and TfR (residues 122–756), nine water molecules, three Ca²⁺ ions, one glycerol molecule and three N -acetyl glucosamine (NAG) moieties. No electron density is observed for the 6x-His tag and N-terminal linker region of TfR, TfR Arg 121, the four C-terminal residues of TfR, the three N-terminal residues of HFE, any of the three potential N -linked carbohydrates in HFE, or two of the three potential carbohydrates in TfR. One NAG residue was modelled at TfR Asn 317. The Ca²⁺ ion in each TfR monomer binds in the same position as one of three samarium ions in uncomplexed TfR¹⁵. The other two samarium sites were unoccupied.

Figures were generated using Molscript, Raster 3D and Setor^{43–45}.

Received 27 August; accepted 21 October 1999.

- Andrews, N. C. & Levy, J. E. Iron is hot: An update on the pathophysiology of hemochromatosis. *Blood* **92**, 1845–1851 (1998).
- Bacon, B. R. *et al.* Molecular medicine and hemochromatosis: at the crossroads. *Gastroenterology* **116**, 193–207 (1999).

- Feder, J. N. *et al.* A novel MHC class I-like gene is mutated in patients with hereditary hemochromatosis. *Nature Genet.* **13**, 399–408 (1996).
- Garcia, K. C., Teyton, L. & Wilson, I. A. Structural basis of T cell recognition. *Annu. Rev. Immunol.* **17**, 369–397 (1999).
- Lebrón, J. A. *et al.* Crystal structure of the hemochromatosis protein HFE and characterization of its interaction with transferrin receptor. *Cell* **95**, 111–123 (1998).
- Feder, J. N. *et al.* The hemochromatosis founder mutation in HLA-H disrupts $\beta 2$ -microglobulin interaction and cell surface expression. *J. Biol. Chem.* **272**, 14025–14028 (1997).
- Waheed, A. *et al.* Hereditary hemochromatosis: effects of C282Y and H63D mutations on association with $\beta 2$ -microglobulin, intracellular processing, and cell surface expression of the HFE protein in COS-7 cells. *Proc. Natl Acad. Sci. USA* **94**, 12384–12389 (1997).
- Feder, J. N. *et al.* The hemochromatosis gene product complexes with the transferrin receptor and lowers its affinity for ligand binding. *Proc. Natl Acad. Sci. USA* **95**, 1472–1477 (1998).
- Parkila, S. *et al.* Association of the transferrin receptor in human placenta with HFE, the protein defective in hereditary hemochromatosis. *Proc. Natl Acad. Sci. USA* **94**, 13198–13202 (1997).
- Aisen, P., Wessling-Resnick, M. & Leibold, E. A. Iron metabolism. *Curr. Opin. Chem. Biol.* **3**, 200–206 (1999).
- Richardson, D. R. & Ponka, P. The molecular mechanisms of the metabolism and transport of iron in normal and neoplastic cells. *Biochim. Biophys. Acta* **1331**, 1–40 (1997).
- Lebrón, J. A. & Bjorkman, P. J. The transferrin receptor binding site on HFE, the Class I MHC-related protein mutated in hereditary hemochromatosis. *J. Mol. Biol.* **289**, 1109–1118 (1999).
- Lebrón, J. A., West, A. P. & Bjorkman, P. J. The hemochromatosis protein HFE competes with transferrin for binding to the transferrin receptor. *J. Mol. Biol.* **294**, 239–245 (1999).
- Gross, C. N., Irinkin, A., Feder, J. N. & Enns, C. A. Co-trafficking of HFE, a nonclassical major histocompatibility complex class I protein, with the transferrin receptor implies a role in intracellular iron regulation. *J. Biol. Chem.* **273**, 22068–22074 (1998).
- Lawrence, C. M. *et al.* Structure of the ectodomain of human transferrin receptor. *Science* **286**, 779–782 (1999).
- Fuchs, H. *et al.* Structural model of phospholipid-reconstituted human transferrin receptor derived by electron microscopy. *Structure* **6**, 1235–1243 (1998).
- Crichton, T. E. *Proteins: Structures and Molecular Properties* (W. H. Freeman and Co., New York, 1993).
- Jones, S. & Thornton, J. M. Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA* **93**, 13–20 (1996).
- Wolf, E., Kim, P. S. & Berger, B. MultiCoil: A program for predicting two- and three-stranded coiled coils. *Protein Sci.* **6**, 1179–1189 (1997).
- Fujinaga, M. *et al.* Refined crystal structure of the seryl-tRNA synthetase from *Thermus thermophilus* at 2.5 Å resolution. *J. Mol. Biol.* **234**, 222–233 (1993).
- Barton, J. C., Sawada-Hirai, R., Rothenberg, B. E. & Acton, R. T. Two novel missense mutations of the HFE gene (I105T and G93R) and identification of the S65C mutation in Alabama hemochromatosis probands. *Blood Cells, Molecules, and Diseases* **25**, 147–155 (1999).
- Egan, T. J., Zak, O. & Aisen, P. The anion requirement for iron release from transferrin is preserved in the receptor–transferrin complex. *Biochemistry* **32**, 8162–8167 (1993).
- Turkewitz, A. P., Schwartz, A. L. & Harrison, S. C. A pH-dependent reversible conformational transition of the human transferrin receptor leads to self-association. *J. Biol. Chem.* **263**, 16309–16315 (1988).
- Bali, P. K., Zak, O. & Aisen, P. A new role for the transferrin receptor in the release of iron from transferrin. *Biochemistry* **30**, 324–328 (1991).
- Bairoch, A. & Apweiler, R. The Swiss-Prot protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* **27**, 49–54 (1999).
- Sipe, D. M. & Murphy, R. F. Binding to cellular receptor results in increased iron release from transferrin at mildly acidic pH. *J. Biol. Chem.* **266**, 8002–8007 (1991).
- Buchegger, F. *et al.* Functional analysis of human/chicken transferrin receptor chimeras indicates that the carboxy-terminal region is important for ligand binding. *Eur. J. Biochem.* **235**, 9–17 (1996).
- Dubljavic, V., Sali, A. & Goding, J. W. A conserved RGD (Arg-Gly-Asp) motif in the transferrin receptor is required for binding to transferrin. *Biochem. J.* **341**, 11–14 (1999).
- Anderson, B. F. *et al.* Structure of human lactoferrin: crystallographic structure analysis and refinement at 2.8 Å resolution. *J. Mol. Biol.* **209**, 711–734 (1989).
- Enns, C. A. & Sussman, H. H. Physical characterization of the transferrin receptor in human placenta. *J. Biol. Chem.* **256**, 9820–9823 (1981).
- Henry, J. B. *Clinical Diagnosis and Management by Laboratory Methods* (W. B. Saunders Co., Philadelphia, 1991).
- Roy, C. N., Penny, D. M., Feder, J. N. & Enns, C. A. The hereditary hemochromatosis protein HFE specifically regulates Tf-mediated iron uptake in HeLa cells. *J. Biol. Chem.* **274**, 9022–9028 (1999).
- Parkila, S. *et al.* Immunohistochemistry of HLA-H, the protein defective in patients with hereditary hemochromatosis, reveals unique pattern of expression in gastrointestinal tract. *Proc. Natl Acad. Sci. USA* **94**, 2534–2539 (1997).
- Waheed, A. *et al.* Association of HFE protein with transferrin receptor in crypt enterocytes of human duodenum. *Proc. Natl Acad. Sci. USA* **96**, 1579–1584 (1999).
- Wilson, I. A. & Bjorkman, P. J. Unusual MHC-like molecules: CD1, Fc receptor, the hemochromatosis gene product, and viral homologs. *Curr. Opin. Immunol.* **10**, 67–73 (1998).
- Verland, S. *et al.* Specific molecular interaction between the insulin receptor and a D product of MHC class I. *J. Immunol.* **143**, 945–951 (1989).
- Stagsted, J. *et al.* Regulation of insulin receptor functions by a peptide derived from a major histocompatibility complex class I antigen. *Cell* **62**, 297–307 (1990).
- Otwinski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- De La Fortelle, E. & Bricogne, G. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276**, 472–494 (1997).
- Kleywegt, G. J. & Jones, T. A. Template convolution to enhance or detect structural features in macromolecular electron density maps. *Acta Crystallogr. D* **53**, 179–185 (1997).
- Collaborative Computational Project No. 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).

42. Brünger, A. T. *et al.* Crystallography and NMR system: A new software system for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
43. Kraulis, P. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
44. Merritt, E. A. & Murphy, M. E. P. Raster3D Version 2.0—a program for photorealistic molecular graphics. *Acta Crystallogr. D* **50**, 869–873 (1994).
45. Evans, S. V. SETOR: hardware lighted three-dimensional solid model representations of macromolecules. *J. Mol. Graphics* **4**, 134–138 (1993).
46. Kawabata, H. *et al.* Molecular cloning of transferrin receptor 2. *J. Biol. Chem.* **274**, 20826–20832 (1999).
47. Nicholls, A., Bharadwaj, R. & Honig, B. GRASP—graphical representation and analysis of surface properties. *Biophys. J.* **64**, A166 (1993).

Acknowledgements

We thank C. M. Lawrence and S. C. Harrison for sharing structural information and coordinates before publication; Z. A. Hamburger for assistance with synchrotron data collection; A. Cohen and M. Soltis for synchrotron support; P. M. Snow and I. Nangiana for expression of TfR; S. Jones for assistance with interface analyses; and A. J. Chirino, C. Enns and J. N. Feder for helpful discussions. M.J.B. was supported by a Cancer Research Institute Postdoctoral fellowship.

Correspondence and requests for materials should be addressed to P.J.B. (e-mail: bjorkman@cco.caltech.edu). Atomic coordinates have been deposited with the PDB under access code 1DE4.

Adaptive-optics corrections available for the whole sky

Roberto Ragazzoni^{*}, Enrico Marchetti[†] & Gianpaolo Valente[‡]

^{*} Astronomical Observatory of Padova vicolo dell'Osservatorio 5, I-35122 Padova, Italy

[†] Centro Galileo Galilei, Apartado de Correos 565, E-38700 Santa Cruz de la Palma, Spain

[‡] Department of Physics and INFN (sez. PD), University of Padova, via F. Marzolo 8, I-35131 Padova, Italy

Adaptive-optics systems can in principle allow a telescope to achieve performance at its theoretical maximum (limited only by diffraction), by correcting in real time for the distortion of starlight by atmospheric turbulence¹. For such a system installed on an 8-m-class telescope^{2,3}, the spatial resolution and sensitivity could be up to 100 times better than conventional imaging^{4,5}. Adaptive-optics corrections have hitherto been achieved only for regions of the sky within a few arcseconds of a bright reference source. But it has been proposed theoretically that by using multiple guide stars, the tomography of atmospheric turbulence could be probed and used to extend adaptive-optics corrections to the whole sky^{6,7}. Here we report the experimental verification of such tomographic⁸ corrections, using three off-axis reference stars ~ 15 arcsec from the central star. We used the observations of the off-axis stars to calculate the deformations of the wavefront of the central star, and then compare them with the real measured values. This tomographic approach is found to reduce variations in the wavefront by $\sim 92\%$. Our result demonstrates that a serious barrier to achieving diffraction-limited seeing over the whole sky has been removed.

On 27 March 1999 from UT 5:42 to UT 6:00, 130 frames of four stars in Aquila ($\alpha = 19^{\text{h}} 29.1^{\text{m}}$, $\delta = +2^{\circ} 39'$) were collected using

the Rotator/Adapter acquisition probe optics⁹ of the 3.6 m Telescopio Nazionale Galileo^{10,11} (TNG).

These four stars lie in an unusual configuration very suitable for our purposes (Fig. 1a). Three stars, ranging from $V \approx 9$ to $V \approx 11$, are arranged at the vertices of a nearly equilateral triangle with a fourth $V \approx 12$ star in its centre. The average distance of the three surrounding stars from the central one is about 15.4 arcsec (ref. 12). The frames have been collected with 0.1 seconds of exposure time on a Peltier-cooled CCD (charge-coupled device; ref. 13), deliberately introducing enough defocus to obtain an extra-focal pupil image for each star approximately 6 arcsec in diameter (Fig. 1b). Seeing was estimated to be 1.2 arcsec at the time of exposures. The central wavelength of observations was 700 nm. Since the delay between two subsequent exposures was of the order of 8 seconds, it is reasonable to assume that the turbulence distorting each image becomes totally uncorrelated after such a delay. During the 18 minutes needed to collect the whole set of frames, no evidence of a significant seeing variation was recorded. Data were collected after a whole night of observations during which thermalization between the telescope and the external environment was ensured. Elevation of the target was roughly 45° .

Those data were used to retrieve wavefront maps for each of the four stars and for any collected image (Fig. 1c). Wavefront maps show the amount of distortion in the optical path for any position in the pupil of the telescope. Instead of carrying out further calculations using the wavefront map as a whole, it is useful to express the latter as a linear combination of a defined set of functions. A suitable set is represented by Zernike polynomials¹⁴. The first ones convey a mathematical description of common Seidel aberrations (such as defocus, coma and astigmatism). These polynomials are conventionally ordered such that the higher their order, the weaker their statistical weight. Thus, for practical purposes, a wavefront map is described by a finite, sufficiently large, number of Zernike polynomials. We fitted each wavefront data with p classical Zernike polynomials up to a complete radial degree order (for instance $p = 27$ polynomials up to 6th radial degree, because the first term, that is the piston, is missing).

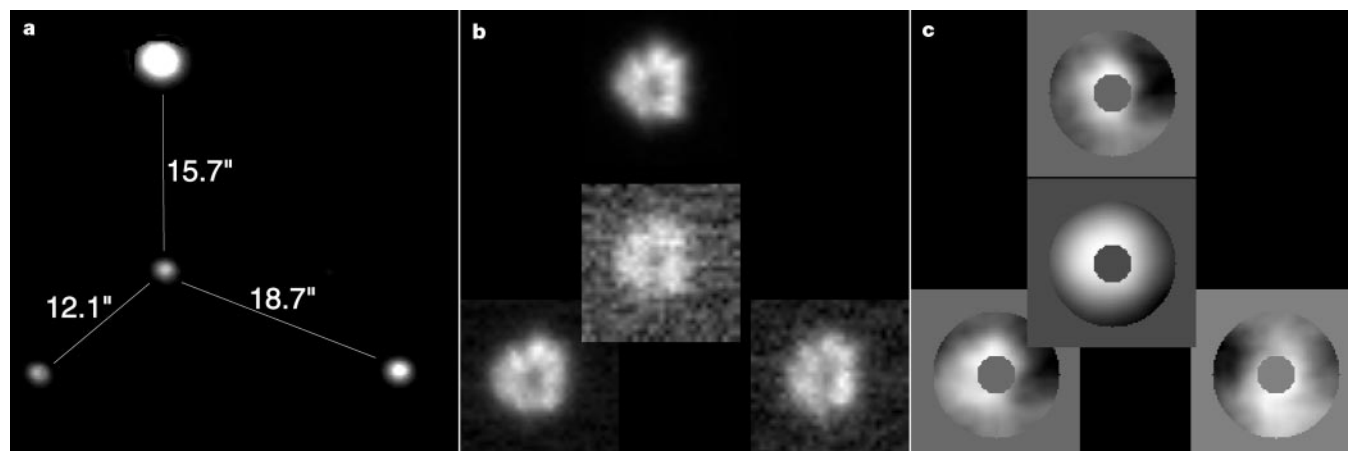


Figure 1 The four observed stars, an example of raw data and the corresponding wavefront maps. **a**, The four observed reference stars. The frames of the four defocused stars have been nominally corrected for detector sensitivity inhomogeneities. **b**, The four pupils have been extracted and stored in four data-cubes of $32 \times 32 \times 130$ pixels each. **c**, Any single-pupil image has been used to retrieve wavefront maps^{21,22}, making use of a Fourier approach²³ to solve the Poisson equation. The wavefront of the central star and the differences between it and the other three are displayed. The starlight-brightness distribution on a defocused pupil depends on both the incoming wavefront curvature and the scintillation. The latter contribution may be cancelled out if, as usual, two extrafocal images are employed to retrieve the wavefront phase. Hickson²¹ suggested using only one defocused pupil to reconstruct the wavefront shape, thus giving also an experimental

demonstration of the technique feasibility²². Nevertheless, those results are still affected by scintillation, which is proportional to the wavefront curvature of each perturbing layer, weighted by its distance from the telescope pupil. In fact, as suggested by Ribak²⁴, such dependence can be used to disentangle, to a limited extent, the origin of the wavefront deformation. The final brightness distribution on the single defocused pupil is given by a linear combination of the wavefront perturbation contributions for each single layer. The conclusions reported here may be applied to all four reference stars, since the effect acts in the same manner. It is also worth noting that the scintillation effects slightly enhance the reliability of our measurements, because of a larger weighting on the high-altitude layers, which are most responsible for the decorrelation among the four reference stars.

It is reasonable that the integration time used (the minimum allowed by our camera) is significantly larger than the atmospheric turbulence correlation time. It means that only the features of a size larger than a characteristic length scale (given by the product of the exposure time and the wind speed) are retained, while the smaller size ones are attenuated. Thus, fitting the measured data with high-order Zernike polynomials is useless for our purposes. During our observing run, ground wind speed was 3 m s^{-1} , leading to a minimum length scale of $\sim 0.3 \text{ m}$. Lacking precise wind-speed measurements at high altitudes, we chose to limit the fitted Zernike polynomials to the first 27.

Each star wavefront map is the result of the bidimensional projection along its direction of all the starlight deformations occurring on several layers at different altitudes. In principle, by combining the information from different stars the three-dimensional distribution of such turbulence can be retrieved. Thus, the turbulence bidimensional projection and the corresponding wavefront along any given line of sight can be estimated, regardless of the existence of a suitable reference star (Fig. 2). Expanding the various wavefront measurements in terms of Zernike polynomials, such an estimate can be described in a linear and compact form⁸ by a modal tomographic matrix M .

Let us call W_T the Zernike polynomial expansion of the beam in the direction of the object of interest, represented by the fourth star at the centre of the quasi-equilateral triangle. Let L be the $3p$ -dimensional vector of the p Zernike polynomial coefficients L_i , $1 \leq i \leq 3$, for each i th star at each vertex of the triangle. Modal tomography assumes the validity of the following relation:

$$W_T = M \cdot L = M \cdot \begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix} \quad (1)$$

Matrix M has been computed row by row using the singular-value decomposition method. In the calculation we omit the piston term. In order to compute the best modal tomographic matrix M , we have maximized the achievable signal-to-noise ratio, using the whole data set.

The final results, displayed in Fig. 3, show that the subtraction of the average of the neighbouring stars (a slight enhancement to the conventional approach) reduces the variance of the single Zernike polynomials by a factor of 5.3 ± 3.3 , while the best M found through the tomographic approach enables a variance reduction by a factor 16.8 ± 4.2 . The large reduction of variance obtained with the simple average may be easily explained as consequence of a significant ground layer.

This experiment proves the feasibility of wavefront sensing in any given direction in the sky, using few off-axis reference stars. The latter can be at a distance much larger than the isoplanatic patch size, whose tiny dimension prevented the use of adaptive optics for whole-sky purposes.

The modal tomography matrix M depends upon the number, altitude and relative strength of the perturbing layers. All these parameters evolve with time and may significantly change during a single observation night. Hence the best matrix M , that is the one producing the best possible tomographic estimate, varies with time. The typical lifetime of M and the time interval to be used for its best estimation are new problems which need to be carefully addressed for the realization of tomographic adaptive optics systems. Even if restriction to a data subset could have led to the investigation of these interesting properties, in our case the augmented noise resulting from the averaging of a smaller number of independent measurements have hidden such effects.

Atmospheric turbulence distribution in altitude h is encoded in a function, usually denoted by $C_N^2(h)$, describing the importance of the air refractive-index fluctuations, causing starlight distortion. Although $C_N^2(h)$ should be treated as a continuous function, current

views on turbulence arrangement suggest that $C_N^2(h)$ support is restricted to a finite, tiny number of limited regions around specific altitudes. It has been shown on general grounds⁸ that M can be computed from the geometry of the star positions and knowledge of $C_N^2(h)$. An independent measure of such a function has not been carried out during our observation. This quantity could have permitted a comparison between the theoretical value of M and the actual one, obtained as a best estimate from our direct measurements. Nevertheless, our work proves the existence of an effective (in the sense of significant reduction of starlight aberration) matrix M , regardless of its similarity to the theoretical expected one.

Site characterization can thus be seen from a completely new point of view. A site characterized by a limited number of different perturbing layers, for which M is expected to be more effective and stable than in other cases, should be classified, to a certain extent, as a better site, regardless of the characteristics of turbulence seen projected over a single line of sight.

The tomographic approach we tested on natural stars may be applied to artificial references as well; these stars, obtained by the excitation of the natural mesospheric sodium layer through a powerful laser¹⁵, immediately extend whole-sky diffraction-limited capability to telescopes of any size. However, a single artificial star cannot solely be used to close an adaptive optics loop. Such an artificial reference, in fact, forms at a limited altitude of $\sim 90 \text{ km}$ and the sampled atmosphere has a conical rather than cylindrical shape, leading to a wavefront estimation error known as conical anisoplanatism. This error scales with the telescope diameter and the

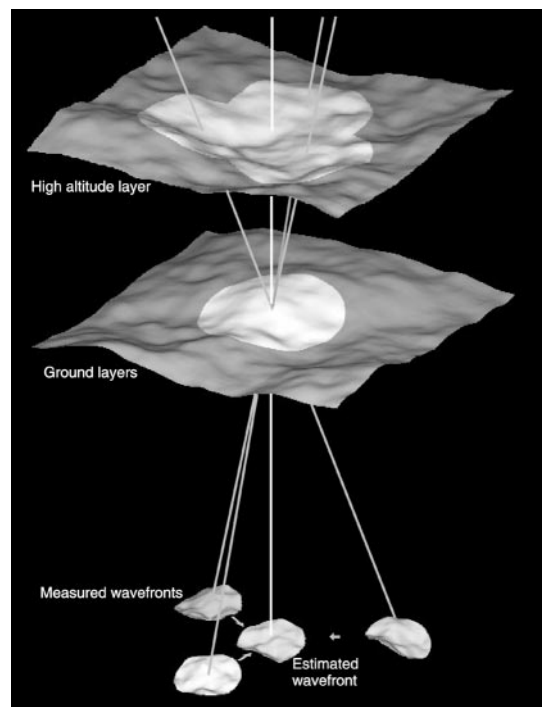


Figure 2 Tomographic measurement of starlight wavefronts exploits turbulence distribution in three dimensions. Light from the source is perturbed by different layers of the Earth's atmosphere. Here we examine two different layers (one is located very close to the ground, as usually happens). For instance, by observing three stars situated at a certain distance from the line of sight of the telescope (the three grey lines), three different starlight wavefront deformations are obtained. None of them can be used to correct on-axis distortion. Tomographic reconstruction of perturbation split into the layers allows us to estimate the wavefront perturbation for a direction (white line) where no reference star is present.

observing wavelength so that laser-guide-star adaptive-optics systems become unusable for 8-m-class telescopes observing in the visible band. Therefore, in order to overcome this problem, multiple laser-guide-stars should be arranged in a constellation.

Scaling up to larger telescope entrance pupils, the limits on the distance of reference stars to the line of sight, in the tomographic approach, are relaxed even further. Next generation 100-m-class telescopes^{16,17} can reach extremely high performance ($V = 38$ with 10 h exposure time, for instance) on condition that they achieve diffraction-limited capabilities. Such a result can be obtained on the whole sky without the need of any artificial bright reference, thus preventing any spurious light pollution¹⁸. Combining this concept with recent ones¹⁹, a limiting magnitude as faint as $V \approx 17$ should be found for a few stars in a patch of several tens of arcmin, in order to provide full turbulence compensation in such an area.

The next big technical challenge is the realization of an adaptive-optics system capable of effectively correcting, in real time, the wavefront approaching from a region of the sky free from bright references. The related concept of multi-conjugate adaptive optics²⁰ is well understood and our measurements, which are essentially open-loop ones, provide convincing proof that this class of systems, which inherently solve the tomographic problem in a closed-loop

fashion, are effective. The efficiency and stability in closed-loop operations are usually much higher than in the open-loop case and, in our opinion, much better results are to be expected than any obtainable from open-loop measurements. Moreover, the technical problems posed by a tomographic approach open up new challenges for wavefront sensing: multiple objects need to be sensed with the aim of possibly detecting wavefront perturbations arising from layers located at different distances from the observer. This is a different problem with respect to the usual wavefront-sensing task in conventional adaptive optics. We think that very soon new classes of wavefront sensors will be realized and tested on the sky.

Such wavefront sensors will be both capable of adapting to the closed-loop situation and compatible with multiple-object wavefront sensing, and may be of use with 8-m-class telescopes leading to whole-sky, diffraction-limited capability, using solely natural stars. □

Received 18 May; accepted 15 October 1999.

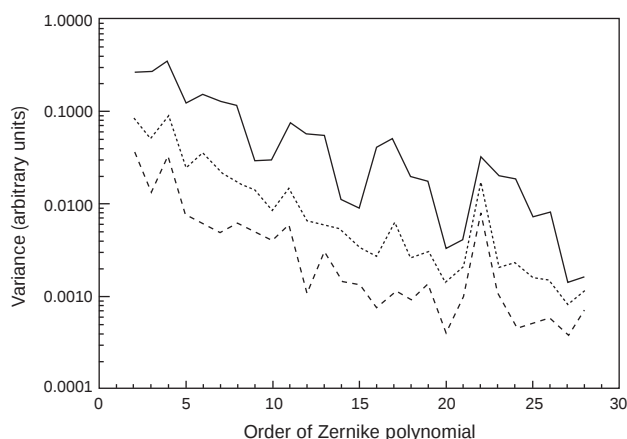


Figure 3 Atmospheric, conventional-technique-corrected and tomographic-corrected turbulence variances. The solid line shows each Zernike polynomial variance as measured on the central star, thus indicating the uncorrected turbulence starlight distortion that prevents diffraction-limited imaging with astronomical telescopes. We computed the residual polynomial variance obtained by subtracting from the central star the simple average of the surrounding ones (dotted line). This represents a somewhat trivial approach to the tomographic problem (actually, in this way no serious attempt at resolution is achieved), although in the special configuration observed, the hope of cancelling out the anti-correlated behaviour of high-altitude effects on the optical beams may be justifiable. This method allowed a 77.4% reductions of wavefront variance compared to that of central star. We compare this result with the average variance reduction determined by the subtraction of Zernike polynomials from each of the three surrounding stars. That procedure allows an average variance reduction of 71.2%. This slightly worse result may be interpreted as further evidence that measurements that are able to rule out the highest-altitude-layer contribution allow a partial correction based on a larger equivalent isoplanatic sky patch size. Such an outcome is in agreement with theoretical suggestions²⁵ and experimental evidence from Rayleigh-based laser guide star data²⁶. The result shown in the plot is hence the best obtainable from conventional adaptive optics when a bright nearby star is not very close to the observed target. The dashed line indicates the residual polynomial variance computed by subtracting from the central star measurements the neighbouring star data, processed through model tomography matrix M . Note that the modal tomography concept allows an additional substantial improvement, reducing the uncorrected wavefront variance by 92.3%. In contrast to the previous case, there seems to be no evident theoretical limit to extending this technique in order to cancel out almost all starlight perturbation.

- Beckers, J. M. Adaptive optics for astronomy—principles, performance, and applications. *Annu. Rev. Astron. Astrophys.* **31**, 13–62 (1993).
- Giacconi, R. et al. Science verification of the Unit Telescope 1 of the ESP Very Large Telescope. *Astron. Astrophys.* **343**, L1–L4 (1999).
- Mountain, C. M., Gillet, F. C. & Oschmann, J. M. Gemini 8-m telescopes. *Proc. SPIE* **3352**, 2–13 (1998).
- Tyson, R. K. *Principles of Adaptive Optics* (Academic, San Diego, 1991).
- Roddier, F. Maximum gain and efficiency of adaptive optics systems. *Publ. Astron. Soc. Pacif.* **110**, 837–840 (1998).
- Murphy, D. V., Primmermann, C. A., Zollars, B. G. & Barclay, H. T. Experimental demonstration of atmospheric compensation using multiple synthetic beacon. *Opt. Lett.* **16**, 1797–1799 (1991).
- Tallon, M. & Foy, R. Adaptive telescope with laser probe—Isoplanatism and cone effect. *Astron. Astrophys.* **235**, 549–557 (1990).
- Ragazzoni, R., Marchetti, E. & Rigaut, F. Modal tomography for adaptive optics. *Astron. Astrophys.* **342**, L53–L56 (1999).
- Ragazzoni, R. Optics for the TNG Rotator/Adapter. *Telescopio Nazionale Galileo Tech. Rep.* **39** 1–8 (1994).
- Barbieri, C. The Galileo Italian National telescope and its instrumentation. *Proc. SPIE* **2871**, 244–255 (1997).
- Bortoletto, F. et al. Commissioning the Italian National Telescope Galileo. *Proc. SPIE* **3352**, 91–101 (1998).
- Sinnott, R. W. Hunting for equilateral triple stars. *Sky Telesc.* March, 100–101 (1999).
- Bortoletto, F. et al. CCD cameras for the Italian national telescope Galileo. *Proc. SPIE* **2654**, 248–258 (1996).
- Noll, R. J. Zernike polynomials and atmospheric turbulence. *J. Opt. Soc. Am.* **66**, 207–211 (1976).
- Happer, W., MacDonald, G. J., Max, C. E. & Dyson, F. J. Atmospheric-turbulence compensation by resonant optical backscattering from the sodium layer in the upper atmosphere. *J. Opt. Soc. Am.* **11-A**, 263–276 (1994).
- Gilmozzi, R. et al. Future of filled aperture telescopes: is a 100-m feasible? *Proc. SPIE* **3352**, 778–791 (1998).
- Mountain, C. M. What is beyond the current generation of groundbased 8-m to 10-m class telescopes and the VLT-I? *Proc. SPIE* **2871**, 579–606 (1997).
- Ragazzoni, R. No laser guide stars for adaptive optics in giant telescopes? *Astron. Astrophys. (Suppl.)* **136**, 205–209 (1999).
- Ragazzoni, R. & Farinato, J. Sensitivity of a pyramidal wave front sensor in closed loop adaptive optics. *Astron. Astrophys.* **350**, L23–L26 (1999).
- Beckers, J. M. Detailed compensation of atmospheric seeing using multiconjugate adaptive optics. *Proc. SPIE* **114**, 215–217 (1989).
- Hickson, P. Wave-front curvature sensing from a single defocused image. *J. Opt. Soc. Am. A* **11**, 1667–1673 (1994).
- Hickson, P. Single-image wavefront curvature sensing. *Proc. SPIE* **2201**, 549–554 (1994).
- Roddier, F. & Roddier, C. Wavefront reconstruction using iterative Fourier transforms. *Appl. Opt.* **30**, 1325–1327 (1991).
- Ribak, E. N. Atmospheric scintillations for measuring remote wave fronts. *Proc. SPIE* **2580**, 166–172 (1995).
- Ragazzoni, R. & Marchetti, E. Rayleigh vs. Sodium beacon for large FoV partial correction. *Atmos. Ocean. Opt.* **8**, 174–176 (1995).
- Christou, J. C., Ellerbroek, B., Fugate, R. Q., Bonaccini, D. & Stanga, R. Rayleigh beacon adaptive optics imaging of ADS 9731: Measurements of the isoplanatic field of view. *Astrophys. J.* **450**, 369–379 (1995).

Acknowledgements

We thank the Telescopio Nazionale Galileo commissioning team, led by F. Bortoletto, including C. Bonoli, A. Cavazza, L. Corcione, M. D'Alessandro, D. Fantinel, D. Gardiol, A. Ghedina, E. Giro, J. C. Guerra Ramón, D. Mancini, J. L. Medina, F. Paulli, C. Pernechele, P. Schipani, G. Tessicini, G. Tranco Lemes, C. Vuerli and A. Zacchei for their invaluable work. Special thanks are also due to L. Salvadori.

Correspondence and requests for materials should be addressed to R.R. (e-mail: ragazzoni@pd.astro.it).

Infrared spectrum of an extremely cool white-dwarf star

S. T. Hodgkin^{*†}, B. R. Oppenheimer[‡], N. C. Hambly[§], R. F. Jameson^{*}, S. J. Smartt^{||†} & I. A. Steele[¶]

^{*} Department of Physics and Astronomy, University of Leicester, University Road, Leicester LE1 7RH, UK

[‡] Department of Astronomy, University of California-Berkeley, 601 Campbell Hall, Berkeley, California 94720, USA

[§] Institute for Astronomy, University of Edinburgh, Blackford Hill, Edinburgh EH9 3HJ, UK

^{||} Isaac Newton Group of Telescopes, Apartado de Correos 321, 38770 Santa Cruz de La Palma, La Palma, Islas Canarias, Spain

[¶] Astrophysics Research Institute, Liverpool John Moores University, Twelve Quays House, Egerton Wharf, Birkenhead L41 1LD, UK

[†] Institute of Astronomy, Cambridge University, Madingley Road, Cambridge CB3 0HA, UK

White dwarfs are the remnant cores of stars that initially had masses of less than 8 solar masses. They cool gradually over billions of years, and have been suggested^{1,2} to make up much of the 'dark matter' in the halo of the Milky Way. But extremely cool white dwarfs have proved difficult to detect, owing to both their faintness and their anticipated similarity in colour to other classes of dwarf stars. Recent improved models³⁻⁵ indicate that white dwarfs are much more blue than previously supposed, suggesting that the earlier searches may have been looking for the wrong kinds of objects. Here we report an infrared spectrum of an extremely cool white dwarf that is consistent with the new models. We determine the star's temperature to be $3,500 \pm 200$ K, making it the coolest known white dwarf. The kinematics of this star indicate that it is in the halo of the Milky Way, and the density of such objects implied by the serendipitous discovery of this star is consistent with white dwarfs dominating the dark matter in the halo.

While white dwarfs may be the predominant component of the dark matter, they can also constrain the age of the Galaxy. The

temperature below which the oldest white dwarfs have not yet had time to cool leads to a cut-off in the luminosity function of these objects. This cut-off has been observed⁶ using white dwarfs drawn from the Luyten Half-Second Catalogue⁷. The turnover at visual magnitude $M_V = 16$, or effective temperature $T_e \approx 4,000$ K, implies an age of 6.5–11 Gyr for the Galactic disk⁸⁻¹⁰. However, if this sample is incomplete, then these ages are underestimates¹¹. Indeed, more recent surveys¹²⁻¹⁴ indicate a density of cool white dwarfs (with temperatures of about 4,000 K) that is higher by a factor of ~ 5 . Early studies of the first luminosity function⁶ and the Hubble Deep Field survey¹⁵⁻¹⁷ showed little evidence for a substantial population of extremely cool (that is, Galactic halo) white dwarfs^{9,18,19}. The new models³⁻⁵ show that these surveys, which looked for red objects, excluded cool white dwarfs.

Two independent models of white-dwarf atmospheres³⁻⁵ stress the dramatic effect of collision-induced absorption by molecular hydrogen on the spectra of very cool, hydrogen-rich white dwarfs. At effective temperatures below 4,000 K, H_2 molecules become abundant in the atmosphere, and, as the collision-induced absorption bands deepen, the peak of the resultant energy distribution shifts to the blue.

A low-luminosity, cool, white dwarf in Taurus, WD0346+246, was serendipitously discovered²⁰ from its large proper motion (1.27 ± 0.04 arcsec yr⁻¹) revealed in UK Schmidt Telescope photographic plates. Initial optical photometry and spectroscopy suggested a temperature $\leq 4,000$ K. We have subsequently obtained infrared photometry and spectra (Fig. 1) and the parallax is now known: 36 ± 5 mas, or a distance of $d = 28 \pm 4$ pc (ref. 21).

The spectral energy distribution presented in Fig. 1 certainly fulfils the expectations of the new theories³⁻⁵, though no model spectra at this low a T_e have yet been reported to enable detailed comparison. (We note that another white dwarf, LHS3250, has recently been observed which shows strong H_2 collision-induced absorption opacity²².)

In the absence of a suitable theoretical model, we can use simple arguments to determine T_e , the luminosity, L , and the radius, r , of the white dwarf, before estimating its mass and age. We fit the optical (wavelengths 0.4–0.8 μ m) spectrum with a black body, and derive a temperature of $3,850 \pm 100$ K. This temperature is solely a parameter used to constrain the actual T_e , the mean radiative

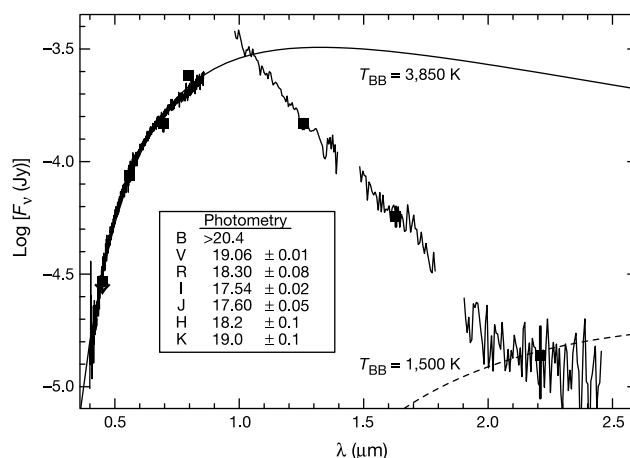


Figure 1 Spectrum of the cool white dwarf WD0346+246. Symbols with error bars are the photometric data from the Jacobus Kapteyn Telescope (JKT) (optical)²⁰ and the UK Infrared Telescope (UKIRT) (infrared), obtained in November 1998. The B-band is an upper limit. The optical spectrum is from ref. 20. We obtained the infrared spectrum with the NIRC²⁰ instrument at the W. M. Keck I Observatory (Hawaii) in February 1999. It covers 1–2.5 μ m at a resolution of $\lambda/\Delta\lambda \approx 150$. The Keck photometric measurements (not shown) agree with the UKIRT values given here. Two fits to the data obtained with a blackbody model are overlaid at $T_{BB} = 3,850$ K (solid curve) and $T_{BB} = 1,500$ K (dashed

curve). This large difference in brightness temperature over the wavelength range plotted demonstrates the effects of H_2 collision-induced absorption in the infrared. We have no explanation for the apparent excess of flux above the blackbody value at around 1 μ m, but we note that models of H_2 –He collision-induced absorption²³ show a reduction in opacity at wavelengths just short of 1 μ m for atmospheres cooler than 4,000 K, and that synthetic spectra⁵ show an excess even for white dwarfs with effective temperatures of about 4,000 K. However, uncertainties in the calibration of the spectra may play a role.

temperature, defined by $L = 4\pi r^2 \sigma T_e^4$, where σ is the Stefan–Boltzmann constant. This blackbody fit is expected in the optical, as H^- is the dominant opacity source here and because it has a very weak wavelength dependence. We can then write:

$$f_{0.4-0.8} = \frac{r^2}{d^2} F_{0.4-0.8}(T = 3,850 \text{ K}) \quad (1)$$

where f is the observed flux, F the blackbody flux and d the distance. Our fitting technique constrains r/d to a precision of about 10%. For a white-dwarf atmosphere with a temperature of $<4,000 \text{ K}$, the H_2 collision-induced absorption opacity increases by almost two orders of magnitude over the wavelength range $1\text{--}2 \mu\text{m}$; at greater wavelengths it flattens²³. So we once again expect a blackbody spectrum at wavelengths above $2 \mu\text{m}$. Knowing r/d and the $2.2\text{-}\mu\text{m}$ flux, we find the temperature of this black body to be $1,500 \text{ K}$. This can be used to determine the contribution to the bolometric flux by the long-wavelength spectrum up to $5 \mu\text{m}$. Beyond $5 \mu\text{m}$, collision-induced absorption opacity becomes less important, and the spectrum will once again be dominated by the hotter and deeper regions of the star's atmosphere. We find the bolometric flux, f_{bol} , by integrating the flux from 0.4 to $2.5 \mu\text{m}$ and adding blackbody components for the regions outside our spectral coverage (we use a linear interpolation over the gap in the spectrum from 0.8 to $1 \mu\text{m}$). These components are a $1,500 \text{ K}$ black body from 2.5 to $5 \mu\text{m}$, and a $3,850 \text{ K}$ black body at wavelengths below $0.4 \mu\text{m}$ and beyond $5 \mu\text{m}$. T_e is derived from f_{bol} by the following:

$$\sigma T_e^4 = f_{\text{bol}} \frac{d^2}{r^2} \quad (2)$$

This gives $T_e = 3,500 \pm 200 \text{ K}$ (minimum formal error). The contributions from the spectral region outside our coverage account for less than 8% of the total flux. Instead of using black bodies to account for this region, we could have linearly extrapolated the data beyond $2.5 \mu\text{m}$. Furthermore, we might reasonably have discounted the excess flux around $1 \mu\text{m}$ (and attributed it to a calibration error). Both of these effects will make only small differences to our estimate of the temperature. Using the distance to WD0346+246 given in ref. 21, we find $L = 7.6(\pm 2.3) \times 10^{28} \text{ erg s}^{-1}$, or $L = 2.0(\pm 0.6) \times 10^{-5} L_\odot$, and $r = 0.012 \pm 0.002 R_\odot$ (here $L_\odot$ and $R_\odot$ are the solar luminosity and radius, respectively). The models²⁴ indicate a mass of $0.65 \pm 0.15 M_\odot$. Therefore WD0346+246 does not have a high mass (like ESO439–26²⁵) and is probably a ‘normal’, albeit unusually cool, white dwarf. The location of WD0346+246 on the white-dwarf sequence in the M_V , $V-I$ diagram also supports this argument²¹. Estimating the age of WD0346+246 is difficult, as we lack important information concerning the atmosphere and core composition. However, the latest cooling models⁴ suggest an age of about 12 Gyr, but with significant uncertainty.

The tangential velocity, v_t , of WD0346+246 provides evidence that it is a member of the Galactic halo. We have not measured a radial velocity for the object, nor are we able to because the spectrum lacks sharp features (such as those from atomic transitions). However, the arguments which follow are insensitive to assumptions about the radial velocity. The proper motion translates to a heliocentric tangential velocity $v_t \approx 170 \text{ km s}^{-1}$. Allowing for the motion of the Sun with respect to the dynamical local standard of rest, v_t can be resolved into a velocity of $\sim 60 \text{ km s}^{-1}$ perpendicular to the Galactic plane (W), and $\sim 150 \text{ km s}^{-1}$ in the Galactic plane directed towards a Galactic longitude, $l = 270^\circ$ (V). Local population I stars in the Galactic disk tend to orbit with velocities similar to the Sun's. The extremely large (V , W) velocity components of WD0346+246 are typical for a population II halo star²⁶.

Is this star simply a white dwarf associated with the known stars in the Galactic halo, or could it be part of a much greater number of white dwarfs accounting for a large fraction—or all—of the halo dark matter suggested by microlensing? The known halo stars have a

mass density of $5.6 \times 10^{-5} M_\odot \text{ pc}^{-3}$ in the solar neighbourhood²⁷. Assuming an initial mass function of slope -2.35 above $1 M_\odot$ and -1.0 below $1 M_\odot$ (ref. 28), the above mass density implies a number density of white dwarfs of $7 \times 10^{-5} \text{ pc}^{-3}$, assuming that all stars more massive than $0.8 M_\odot$ have become white dwarfs. If white dwarfs accounted for half of the measured halo dark matter, then the implied number density would need to be two orders of magnitude higher at $7.9 \times 10^{-3} \text{ pc}^{-3}$ (refs 2, 29). WD0346+246 is half a magnitude brighter than the I-plate limit²⁰, so similar objects could have been detected out to a distance of 35 pc, or a volume of $\sim 100 \text{ pc}^3$ for one $5^\circ \times 5^\circ$ photographic plate from the Schmidt telescope. Thus on one Schmidt plate we expect to see either 0.008 white dwarfs associated with the known Galactic halo stars, or 0.9 ‘dark matter’ white dwarfs. This one serendipitous discovery on one plate clearly favours the latter possibility, but further survey work is needed for a definite conclusion. $\square$

Received 7 June; accepted 15 November 1999.

- Chabrier, G., Segretain, L. & Mera, D. Contribution of brown dwarfs and white dwarfs to recent microlensing observations and the halo mass budget. *Astrophys. J.* **468**, L21–L24 (1996).
- Alcock, C. *et al.* The MACHO project Large Magellanic Cloud microlensing results from the first two years and the nature of the Galactic dark halo. *Astrophys. J.* **486**, 697–726 (1997).
- Hansen, B. M. S. Old and blue white-dwarf stars as a detectable source of microlensing events. *Nature* **394**, 860–862 (1998).
- Hansen, B. M. S. Cooling models for old white dwarfs. *Astrophys. J.* **502**, 680–695 (1999).
- Saumon, D. & Jacobsen, S. B. Pure hydrogen model atmospheres for very cool white dwarfs. *Astrophys. J.* **511**, L107–L110 (1999).
- Liebert, J., Dahn, C. C. & Monet, D. G. The luminosity fraction of white dwarfs. *Astrophys. J.* **332**, 891–909 (1988).
- Luyten, W. J. *NLTT Catalog* (Univ. Minnesota Press, Minneapolis, 1979).
- Winget, D. E. *et al.* An independent method for determining the age of the universe. *Astrophys. J.* **315**, L77–L81 (1987).
- Wood, M. A. Constraints on the age and evolution of the Galaxy from the white dwarf luminosity function. *Astrophys. J.* **386**, 539–561 (1992).
- Leggett, S. K., Ruiz, M. T. & Bergeron, P. The cool white dwarf luminosity function and the age of the Galactic Disk. *Astrophys. J.* **497**, 294–302 (1998).
- Oswalt, T. D., Smith, J. A., Wood, M. A. & Hintzen, P. A lower limit of 9.5 Gyr on the age of the Galactic disk from the oldest white dwarf stars. *Nature* **382**, 692–694 (1996).
- Knox, R. A., Hawkins, M. R. S. & Hambly, N. C. A survey for cool white dwarfs and the age of the Galactic Disk. *Mon. Not. R. Astron. Soc.* **306**, 636–752 (1999).
- Ruiz, M. T. New cool degenerate stars. *Astron. J.* **111**, 1267–1270 (1996).
- Festin, L. The luminosity function of white dwarfs and M dwarfs using dark nebulae as opaque outer screens. *Astron. Astrophys.* **336**, 883–894 (1998).
- Flynn, C., Gould, A. & Bahcall, J. N. Hubble Deep Field constraint on baryonic dark matter. *Astrophys. J.* **466**, L55–L58 (1996).
- Elson, R. A. W., Santiago, B. X. & Gilmore, G. F. Halo stars, starbursts, and distant globular clusters: A survey of unresolved objects in the Hubble Deep Field. *New Astron.* **1**, 1–16 (1996).
- Mendez, R. A., Minniti, D., de Marchi, G., Baker, A. & Couch, W. J. Starcounts in the Hubble Deep Field: Constraining Galactic structure models. *Mon. Not. R. Astron. Soc.* **283**, 666–672 (1996).
- Graff, D. S., Laughlin, G. & Freese, K. MACHOs, white dwarfs and the age of the universe. *Astrophys. J.* **499**, 7–19 (1998).
- Tamanaha, C. M., Silk, J., Wood, M. A. & Winget, D. E. The white dwarf luminosity function—A possible probe of the Galactic Halo. *Astrophys. J.* **358**, 164–169 (1990).
- Hambly, N. C., Smartt, S. J. & Hodgkin, S. T. WD0346+246: A very low luminosity, cool degenerate in Taurus. *Astrophys. J.* **489**, L157–L160 (1997).
- Hambly, N. C. *et al.* On the parallax of WD0346+246: a Halo white dwarf candidate. *Mon. Not. R. Astron. Soc.* **309**, L33–L36 (1999).
- Harris, H. C. *et al.* A very low-luminosity, very cool, DC white dwarf. *Astrophys. J.* **524**, 1000–1007 (1999).
- Borysow, A., Jørgensen, U. G. & Zheng, C. Model atmospheres of cool, low-metallicity stars: the importance of collision-induced absorption. *Astron. Astrophys.* **324**, 185–195 (1997).
- Wood, M. A. in *White Dwarfs* (eds Koester, D. & Werner, K.) 41–45 (Springer, Berlin Heidelberg, 1995).
- Ruiz, M. T., Bergeron, P., Leggett, S. K. & Anguita, C. The extremely low luminosity white dwarf ESO 439–26. *Astrophys. J.* **455**, L159–L162 (1995).
- Ryan, S. G. & Norris, J. E. Subdwarf studies. II—abundances and kinematics from medium resolution spectra. *Astron. J.* **101**, 1835–1864 (1991).
- Dahn, C. C., Liebert, J., Harris, H. C. & Guetter, H. H. in *The Bottom of the Main Sequence—and Beyond* (ed. Tinney, C. G.) 239–248 (Springer, Berlin Heidelberg, 1995).
- Gizis, J. E. & Reid, I. N. M subdwarfs: the population II luminosity function. *Astron. J.* **117**, 508–520 (1999).
- Richer, H. B. *et al.* Isochrones and luminosity functions for old white dwarfs. *Astrophys. J.* (in the press).
- Matthews, K. & Soifer, B. T. The near infrared camera on the W.M. Keck telescope. *Exp. Astron.* **3**, 77–84 (1994).

Acknowledgements

The UK Infrared Telescope is operated by the Joint Astronomy Centre on behalf of the UK Particle Physics and Astronomy Research Council. Some of the data presented here were obtained at the W. M. Keck Observatory, which is operated as a scientific partnership

among the California Institute of Technology, the University of California and NASA. The Observatory was made possible by the support of the W. M. Keck Foundation. S.T.H. and S.J.S. were supported by the PPARC; B.R.O. was supported by FUTDI and a Hubble postdoctoral research fellowship.

Correspondence and requests for materials should be addressed to S.T.H. (e-mail: sth@ast.cam.ac.uk).

Tunnelling between the edges of two lateral quantum Hall systems

W. Kang^{*†}, H. L. Stormer^{†‡}, L. N. Pfeiffer[†], K. W. Baldwin[†] & K. W. West[†]

^{*} James Franck Institute and Department of Physics, University of Chicago, Chicago, Illinois 60637, USA

[†] Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA

[‡] Department of Physics and Department of Applied Physics, Columbia University, New York, New York 1002, USA

The edge of a two-dimensional electron system in a magnetic field consists of one-dimensional channels that arise from the confining electric field at the edge of the system^{1–3}. The crossed electric and magnetic fields cause electrons to drift parallel to the sample boundary, creating a chiral current that travels along the edge in only one direction. In an ideal two-dimensional electron system in the quantum Hall regime, all the current flows along the edge^{4–6}. Quantization of the Hall resistance arises from occupation of N one-dimensional edge channels, each contributing a conductance of e^2/h (refs 7–11). Here we report differential conductance measurements, in the integer quantum Hall regime, of tunnelling between the edges of a pair of two-dimensional electron systems that are separated by an atomically precise, high-quality, tunnel barrier. The resultant interaction between the edge states leads to the formation of new energy gaps and an intriguing dispersion relation for electrons travelling along the barrier: for example, we see a persistent conductance peak at zero bias voltage and an absence of tunnelling features due to electron spin. These features are unexpected and are not consistent with a model of weakly interacting edge states. Remnant disorder along the barrier and charge screening may each play a role, although detailed numerical studies will be required to elucidate these effects.

The one-dimensional edge channels of a two-dimensional electron system (2DES) in the quantum Hall regime represent a prototypical one-dimensional electronic system. Other such one-dimensional model systems include semiconductor-based quantum wires¹², carbon nanotubes^{13,14}, and molecular chain materials¹⁵. While these systems force the electrons to move along a preferential direction, the edge channels of a 2DES in the quantum Hall regime are unique in their ability to adjust themselves spatially as well as energetically. A localized defect is easily avoided by the edge current by simply skirting the impurity potential. In real samples, variations in the potential landscape can generate a complex topology of edge channels^{16,17}. Consequently, most real edge channels are ill defined, both spatially and energetically, and it becomes difficult to generalize various experimental geometries in terms of one-dimensional edge states. Although typical experiments seek to study edge states in lithographically defined geometries^{18,19}, such edges typically fluctuate on the scale of a magnetic length. In order to address the energetics of edge channels, well-defined geometries with clearly delineated edges are essential. For tunnelling experiments in particular, where distance factors exponentially into the tunnelling current, exact knowledge of the shape and the magnitude of the barrier is crucial when comparing experimental results with model calculations.

Our 2DES–barrier–2DES (2D–2D) tunnelling device consists of two stripes of 2DES separated by an atomically precise 88-Å-thick semiconductor barrier, fabricated by cleaved edge overgrowth (Fig. 1a). We explore the lateral tunnelling between these two physically separated 2DESs in the quantum Hall regime, as shown in Fig. 1b. Figure 1c shows the differential conductance of the high-density sample at 9.2 T and 6.0 T. At 9.2 T, dI/dV_{bias} (where I is the current, and V_{bias} is bias voltage) is strongly suppressed around zero bias, while oscillatory conductance peaks appear above a threshold electric field. Each peak represents the onset of an additional tunnelling path through the barrier. Successive conductance peaks are nearly equally spaced, with a spacing of the order of the cyclotron energy $\hbar\omega_c$ (~ 17 meV at 10 T), and the amplitude of their oscillation decreases with increasing V_{bias} . Most of the traces of dI/dV_{bias} resemble the 9.2-T conductance data. But for certain ranges of magnetic field, B , there is no threshold to tunnelling, and a very sharp and intense conductance peak arises at zero bias, as illustrated by the 6-T data in Fig. 1c, lower panel.

Figure 2 shows an image map of the dI/dV_{bias} data as a function of Landau level filling factor, $\nu = nh/eB$ (where n is electron density and e is electron charge) and the normalized bias voltage, $eV_{\text{bias}}/\hbar\omega_c$. Blue and red represent small and large dI/dV_{bias} signals, respectively. With scaled axes, the disparate data from two different samples with

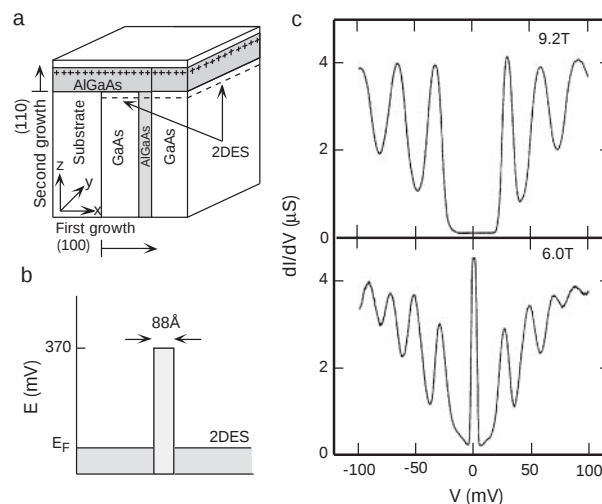


Figure 1 Structure and differential conductance measurements of our 2D–2D tunnelling device. **a**, The junctions are fabricated by cleaved edge overgrowth in molecular beam epitaxy (MBE)^{22,23}. The first growth on a standard (100)GaAs substrate consists of an undoped 13- μm GaAs layer followed by an 88-Å-thick digital alloy of undoped $\text{Al}_{0.1}\text{Ga}_{0.9}\text{As}/\text{AlAs}$, and completed by a 14- μm layer of undoped GaAs. This triple-layer sample is cleaved along the (110) plane in an MBE machine and a modulation-doping sequence is grown over the exposed edge. It consists of a 3,500-Å-thick AlGaAs layer, delta-doped with Si at a distance of 500 Å from the interface. Carriers from the Si impurities transfer only into the GaAs layers of the cleaved edge, forming two stripes of 2DES of widths 13 μm and 14 μm , separated from each other by an 88-Å-thick, 370-meV-high $\text{Al}_{0.1}\text{Ga}_{0.9}\text{As}/\text{AlAs}$ barrier. The sample is fabricated into a mesa incorporating the barrier and the two 2DESs. Contacts are made to the 2DESs, far away from the tunnelling region. **b**, Schematic band structure of the 2D–2D tunnelling device. Two different samples with electron density of $n = 1.1 \times 10^{11} \text{ cm}^{-2}$ and $n = 2.0 \times 10^{11} \text{ cm}^{-2}$ are studied. From a simultaneously grown monitor wafer, we estimate the mobility of the 2DESs in the device to be $\sim 1 \times 10^5 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$. **c**, Two representative traces of differential conductance through the tunnelling barrier with electron density $n = 2.0 \times 10^{11} \text{ cm}^{-2}$. A low-frequency a.c. technique (typically 10 μV amplitude) is employed to measure the differential conductance, dI/dV_{bias} , through the barrier in the presence of a d.c. voltage bias, V_{bias} . The samples are measured at $T = 300$ mK in a magnitude field. Differential conductances, dI/dV_{bias} , nearly vanish near zero bias, while conductance oscillations of the order of $10^{-6} \Omega^{-1}$ are found at high bias.

very different electron densities join, and produce a universal tunnelling map for the 2D-2D tunnelling. An intriguing pattern arises with a continuous progression of conductance maxima (squares and circles in Fig. 2) from the low-density to the high-density data. The dI/dV_{bias} plot exhibits gap-like thresholds that define regions of vanishing dI/dV_{bias} in which the tunnelling is strongly suppressed. Around zero bias, tunnelling is suppressed in a bell-shaped region for filling factors $\nu < 1$. Similar thresholds occur at higher fillings near $\nu \approx 2$. Between fillings $\nu = 1.2-1.5$ and $2.2-2.5$, a sharp conductance peak dominates at zero bias. The positions of the oscillatory conductance peaks vary roughly linearly at high fillings, and produce a fan of maxima that branches out from zero bias. Some of these branches cross each other above $\nu > 1$. For fillings $\nu < 1$, the different branches tend to saturate.

By analogy to the edge-state transport in the quantum Hall 2DES, we consider the 2D-2D tunnelling in terms of one-dimensional edge states in the presence of a barrier. The edge states around the perimeter of the 2D-2D system are the same as in a simple 2DES, whose dispersion relation along the edge is approximately parabolic. This is no longer the case along the barrier. Figure 3a illustrates the spatial dependence of the Landau level energy in its vicinity. The energy levels $E(x)$ in Fig. 3a correspond to the energy of an electron with orbit guiding centre x . As $x = -k_y l_0^2$, where the magnetic length $l_0 = \sqrt{\hbar/eB}$, the energy diagram of Fig. 3a is also the one-dimensional dispersion relation, $E(k_y)$, for electrons travelling along the barrier in the y -direction with velocity $v = \hbar^{-1} dE/dk_y = \hbar^{-1} l_0^2 dE(x)/dx$. Electrons on the left side of the barrier counter-propagate with respect to those on the right side of the barrier. At the intersections of the two sets of rising Landau ladders, there exist two oppositely travelling, degenerate edge states with the same wavevector, $k_y = -x/l_0^2$. The degeneracy at the crossings is lifted by the formation of a series of small energy gaps that separate the symmetric and antisymmetric combinations of the underlying wavefunctions as seen in Fig. 3a. Altogether, Fig. 3a bears a surprising resemblance to the data of Fig. 2.

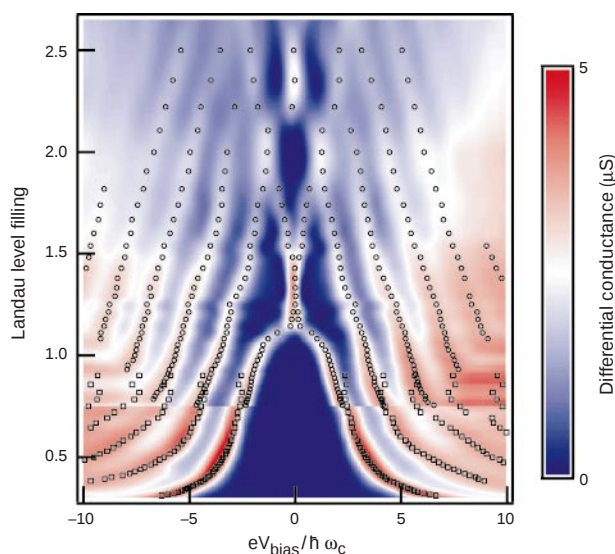


Figure 2 Evolution of the differential conductance of 2D-2D tunnelling devices under high magnetic fields. The Landau level filling factor, $\nu = nh/eB$, and the normalized bias, $eV_{\text{bias}}/\hbar\omega_c$, are used as universal axes for the 2D-2D tunnelling data from samples with electron densities of $n = 2.0 \times 10^{11} \text{ cm}^{-2}$ (Landau level filling > 0.75) and $n = 1.1 \times 10^{11} \text{ cm}^{-2}$ (Landau level filling < 0.75). Their maxima are indicated by circles and squares, respectively. Blue (red) regions represent minimum (maximum) conductance. In the blue region, conductances, dI/dV_{bias} , nearly vanish, while the red areas represent conductances of the order of $10^{-6} \Omega^{-1}$.

Electronic transport along the barrier depends crucially on the position of the Fermi energy with respect to this complex one-dimensional barrier level structure (ref. 20, and S. M. Girvin, personal communication). When the Fermi level is below the first maximum, say at E_1 in Fig. 3a, electrons in both 2DESs follow two separate, counter-propagating tracks along the junction, as depicted in Fig. 3a, left inset. Although travelling along the barrier, they have very different k_y -wave vectors and are practically uncoupled. Therefore, tunnelling through the barrier is negligible, corresponding to the absence of a peak at zero bias at low filling factor.

As the Fermi level reaches the first maximum at E_2 in Fig. 3a, the edge states on both sides of the barrier have identical wavevector and resonate, which leads to substantial tunnelling. In fact, the tracks of the wavefunctions have insignificantly changed from the situation shown in the left inset of Fig. 3a, but coupling between both edges has vastly increased due to the equivalence of their k_y -momentum. Consequently, at this particular position of the Fermi energy E_F , the conductance at zero bias becomes finite, as observed in the zero bias peak of our data. According to the barrier level scheme in Fig. 3a, the edge states from the second Landau level, $N=1$, are also occupied for this range of fillings, but remain uncoupled.

As the Fermi level rises slightly above this critical position and into the gap of the dispersion relation, electrons can no longer travel along the barrier, as shown in the right inset of Fig. 3a. Consequently, coupling between both 2DESs should vanish and tunnelling should cease. This represents an interesting paradox. If electrons can no longer travel along the barrier and, due to the chirality of the edge channel, are not allowed to backscatter, one would think they need to tunnel. This should give rise to a conductance of $\sim e^2/h$, which is, however, two orders of magnitude larger than observed. The detailed tunnelling and scattering model at this position of E_F remains unclear. In any case, our model does

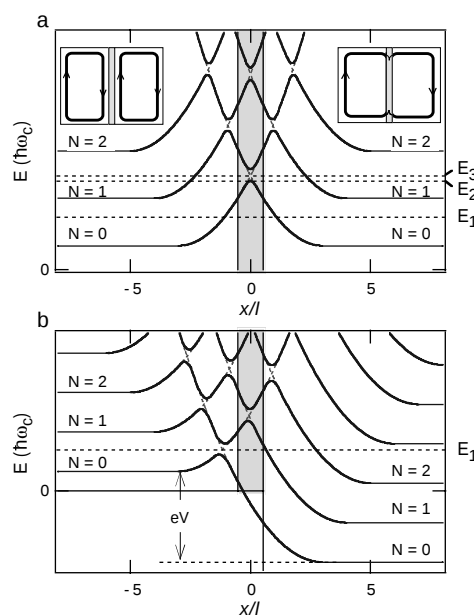


Figure 3 Schematic energy dependence of the Landau levels in the vicinity of the barrier. **a**, For zero voltage bias. Far away from the barrier, Landau levels are equally spaced, $E_N = (N + 1/2)\hbar\omega_c$. As electrons approach the edge of the barrier, their energy rises parabolically, $E/\hbar\omega_c = (|x/l_0| + \sqrt{N+1})^2 + (N+1/2)$, $N=0,1,2,\dots$ and for $|x/l_0| \leq \sqrt{N+1}$, otherwise $E/\hbar\omega_c = N + 1/2$, where $l_0^2 = \hbar/eB$ is the magnetic length. In the vicinity of the barrier, these parabolas anti-cross and create a gapped spectrum. The traces represent the energy of an electron whose guiding centre is located at x . N is the Landau level index and E_1 to E_3 represent Fermi energies at different filling factors. Insets represent the in-plane track of the electronic wavefunction (see text for details). **b**, As **a** but with a bias of $\sim 2\hbar\omega_c$ applied across the barrier.

not seem to agree with experiment, in which the zero-bias peak persists for a range of filling factors $1.2 < \nu < 1.5$, rather than existing at just one magnetic field intensity. This may be related to electron scattering along the barrier, which relaxes k -conservation and washes out the dispersion relation. It is also unclear why the conductance vanishes abruptly at $\nu \approx 1.5$. This may coincide with E_F reaching the top of the energy gap of the $N = 0$ branch, where both 2DESs couple once again along the barrier. A sharp zero-bias conduction and a strong suppression of tunnelling alternate as E_F moves through the higher-lying gaps shown in Fig. 3a.

What is the origin of the complex overall pattern of Fig. 2, away from $V_{\text{bias}} = 0$, and what is its relationship to Fig. 3a? At finite bias, one set of Landau levels of Fig. 3a is raised with respect to the other by an energy eV_{bias} , as shown in Fig. 3b. This results in a shift of the intercepts, and the conditions for onset of conduction are moved to a different energy. For example, when $E_F = E_1$ in Fig. 3a, tunnelling across the barrier is inhibited, owing to the absence of coupling of the edge states. In Fig. 3b the application of eV_{bias} has shifted the crossing so as to coincide with E_1 . This enhances immensely the coupling between both 2DESs, in analogy to position E_2 in Fig. 3a, and provides an explanation for the shift of the onset of tunnelling to higher V_{bias} for smaller ν in Fig. 2.

In general, whenever the Fermi energy coincides with one of the crossing points, electrons can tunnel through the barrier and relax to the Fermi energy of the 2DES, lying eV_{bias} below. Each additional coincidence of the levels produces a peak in dI/dV_{bias} . We can track their positions if we neglect the small gaps at the crossings. The Landau ladders on both sides are approximately described by:

$$\frac{E_L}{\hbar\omega_c} = \left(\frac{x}{l_0} + \sqrt{N_L + 1} \right)^2 + \left(N_L + \frac{1}{2} \right)$$

and

$$\frac{E_R}{\hbar\omega_c} = \left(\frac{x}{l_0} - \sqrt{N_R + 1} \right)^2 + \left(N_R + \frac{1}{2} \right) - \frac{eV_{\text{bias}}}{\hbar\omega_c} \quad (1)$$

Flat Landau levels are found for $|x/l_0| \geq (N + 1)^{0.5}$.

Assuming that N_R and N_L may be neglected in comparison with $eV_{\text{bias}}/2\hbar\omega_c$, which is justified for large sections of Fig. 2, and assuming that $E_L/\hbar\omega_c \approx \nu$, which holds for very broad Landau levels, as is probably the case in our lower-mobility specimen, we arrive at

$$\nu \approx \left(-\frac{eV_{\text{bias}}}{2\hbar\omega_c} + \sqrt{N_L + 1} \right)^2 + (N_L + 1/2) \quad (2)$$

as the condition for coincidence.

Equation (1)—which (together with its left-right mirror image) describes the barrier level structure of Fig. 3a—is identical to equation (2). And equation (2) identifies the peaks in the differential conductance in Fig. 2, as long as $E_L/\hbar\omega_c$ is replaced by ν and x/l_0 is replaced by $eV_{\text{bias}}/2\hbar\omega_c$. This resolves the puzzle of the similarity between both graphs.

Although we can account for the general features of our experimental results, many aspects of the data remain unresolved and require an explanation beyond our simple model. One such feature is the extended existence of the zero-bias peak, which is expected to occur only at the point of coincidence of E_F with the edges of the gaps. The origin of this discrepancy may be a relaxed k -conservation due to remnant disorder along the barrier, which scatters electrons, broadens the dispersion relation and therefore extends the allowed energy range for strong tunnelling. Such a broadening may also account for the conductance which is much reduced compared to e^2/h : the sharp resonance with universal conductance is broadened into a band of much lower, average conductance. However, the sharpness of the zero-bias peak as a function of voltage bias implies

limited broadening. A proper accounting of the role of disorder in 2D-2D tunnelling requires a detailed, quantitative analysis of our 2D-2D device.

Another puzzling feature of our data is the position in filling factor at which the zero-bias conductance peak appears. According to Fig. 3a, the first coincidence of the Landau ladders occurs at $\nu \approx 4$, somewhat above the $N = 1$ Landau level, contrary to the experimental value of $\nu \approx 1.2$. Similarly, the next coincidence is expected for $\nu > 6$, while it is observed at $\nu \approx 2.2$. These observations point to a shifting of the levels in addition to a possible broadening. It could arise from self-consistent screening and from an accumulation of charge in the vicinity of the barrier²¹, which modifies the level scheme.

The influence of the electron spin on our experiment, as well as on the dispersion relation in Fig. 3a, remains unclear. While the Zeeman splitting in GaAs is only $\sim 1/70$ of the cyclotron splitting, partial and spatially dependent occupation of Landau levels near the barrier can produce large spatially dependent enhancement of the g -factor^{1,2}. This can give rise to additional coincidences—possibly over ranges of fillings—and will appear in the tunnelling characteristics, in particular, if spin-flip scattering is strong. This absence of spin-dependent features in the tunnelling data remains a central puzzle. Numerical studies of our simple physical system should provide guidance as to the relative importance of different mechanisms. □

Received 22 July; accepted 7 October 1999.

1. Prange, R. E. & Girvin, S. M. (eds) *The Quantum Hall Effect* 2nd edn (Springer, New York, 1990).
2. Das Sarma, S. & Pinczuk, A. (eds) *Perspectives in Quantum Hall Effects* (Wiley Inter-Science, New York, 1997).
3. Halperin, B. I. Quantized Hall conductance, current-carrying edge states, and the existence of extended states in a two-dimensional disordered potential. *Phys. Rev. B* **25**, 2185–2188 (1983).
4. MacDonald, A. H. & Streda, P. Quantized Hall effect and edge currents. *Phys. Rev. B* **29**, 1616–1619 (1984).
5. Apenko, S. M. & Lozovik, Yu. E. J. The quantized Hall effect in strong magnetic fields. *J. Phys. C* **18**, 1197–1203 (1985).
6. Fontein, P. F. *et al.* Spatial potential distribution in GaAs/Al_xGa_{1-x}As heterostructures under quantum Hall conditions studied with the linear electro-optic effect. *Phys. Rev. B* **43**, 12090–3 (1991).
7. Buttiker, M. Absence of backscattering in the quantum Hall effect in multiprobe conductors. *Phys. Rev. B* **38**, 9375–9389 (1988).
8. Streda, P., Kucera, J. & MacDonald, A. H. Edge states, transmission matrices, and the Hall resistance. *Phys. Rev. Lett.* **59**, 1973–1975 (1987).
9. Jain, J. K. & Kivelson, S. A. Landauer-type formulation of quantum-Hall transport: critical currents and narrow channels. *Phys. Rev. B* **37**, 4276–4279 (1988).
10. Haug, R. J., MacDonald, A. H., Streda, P. & von Klitzing, K. Quantized multichannel magnetotransport through a barrier in two dimensions. *Phys. Rev. Lett.* **61**, 2797–2800 (1988).
11. Washburn, S., Fowler, A. B., Schmid, H. & Kern, D. Quantized Hall effect in the presence of backscattering. *Phys. Rev. Lett.* **61**, 2801–2804 (1988).
12. Yacoby, A. *et al.* Non-universal conductance quantization in quantum wires. *Phys. Rev. Lett.* **77**, 4612–4615 (1996).
13. Wildoer, J. W. G., Venema, L. C., Rinzler, A. G., Smalley, R. E. & Dekker, C. Electronic structure of atomically resolved carbon nanotubes. *Nature* **391**, 59–62 (1998).
14. Odom, T. W., Huang, J., Kim, P. & Lieber, C. M. Atomic structure and electronic properties of single walled carbon nanotubes. *Nature* **391**, 62–64 (1998).
15. Ishiguro, T., Yamaji, K. & Saito, G. *Organic Superconductors* 2nd edn (Springer, New York, 1998).
16. Tessmer, S. H., Glicofridis, P. I., Ashoori, R. C., Levitov, L. S. & Melloch, M. R. Surface charge accumulation imaging of a quantum Hall liquid. *Nature* **392**, 51–54 (1998).
17. McCormick, K. L. *et al.* Scanned potential microscopy of edge and bulk currents in the quantum Hall regime. *Phys. Rev. B* **59**, 4654–4657 (1999).
18. Goldman, V. J. & Su, B. Resonant tunnelling in the quantum Hall regime: measurement of fractional charge. *Science* **267**, 1010–1012 (1995).
19. Tarucha, S., Honda, T. & Saku, T. Reduction of quantized conductance at low temperatures observed in 2 to 10 μm -long quantum wires. *Solid State Commun.* **94**, 413–18 (1995).
20. Ho, T. L. Oscillator tunnelling between quantum Hall systems. *Phys. Rev. B* **50**, 4524–4533 (1994).
21. Wulf, U., Gudmundsson, V. & Gerhardt, R. R. Screening properties of the two-dimensional electron gas in the quantum Hall regime. *Phys. Rev. B* **38**, 4218–4230 (1988).
22. Pfeiffer, L. N. *et al.* Formation of a high quality two-dimensional electron gas on cleaved GaAs. *Appl. Phys. Lett.* **56**, 1697–1699 (1990).
23. Chang, A. M., Pfeiffer, L. N. & West, K. W. Observation of chiral Luttinger behavior in electron tunnelling into fractional quantum Hall edges. *Phys. Rev. Lett.* **77**, 2538–2541 (1996).

Acknowledgements

We thank S. M. Girvin for insight into the energetics of our 2D-2D device geometry. We also thank R. de Picciotto, A. M. Chang, T. L. Ho and J. P. Eisenstein for discussions.

Correspondence and requests for materials should be addressed to W.K. (e-mail: wkang@rainbow.uchicago.edu).

Magnetic trapping of neutrons

P. R. Huffman^{*†}, C. R. Brome^{*}, J. S. Butterworth^{*‡}, K. J. Coakley[§],
M. S. Dewey[†], S. N. Dzhosyuk^{*}, R. Golub^{||}, G. L. Greene[¶], K. Habicht^{||},
S. K. Lamoreaux[¶], C. E. H. Mattoni^{*}, D. N. McKinsey^{*}, F. E. Wietfeldt^{*†}
& J. M. Doyle^{*}

^{*}Harvard University, 17 Oxford Street, Cambridge, Massachusetts 02138, USA

[†]National Institute of Standards and Technology, 100 Bureau Drive, MS 8461, Gaithersburg, Maryland 20899, USA

[§]National Institute of Standards and Technology, 325 Broadway, MS898.02, Boulder, Colorado 80303, USA

^{||}Hahn-Meitner-Institut, Glienicker Strasse 100, D-14109 Berlin, Germany

[¶]University of California, Los Alamos National Laboratory, PO Box 1663, Los Alamos, New Mexico 87545, USA

[‡]Present address: Institut Laue-Langevin, BP 156-6 rue Jules Horowitz, 38042, Grenoble (Cedex 9), France.

Accurate measurement of the lifetime of the neutron (which is unstable to beta decay) is important for understanding the weak nuclear force¹ and the creation of matter during the Big Bang². Previous measurements of the neutron lifetime have mainly been limited by certain systematic errors³; however, these could in principle be avoided by performing measurements on neutrons stored in a magnetic trap. Neutral-particle and charged-particle traps are widely used for studying both composite and elementary particles, because they allow long interaction times and isolation of particles from perturbing environments⁴. Here we report the magnetic trapping of neutrons. The trapping region is filled with superfluid ⁴He, which is used to load neutrons into the trap and as a scintillator to detect their decay. Neutrons in the trap have a lifetime of 750^{+330}_{-200} seconds, mainly limited by their beta decay rather than trap losses. Our experiment verifies theoretical predictions regarding the loading process and magnetic trapping of neutrons. Further refinement of this method should lead to improved precision in the neutron lifetime measurement.

A more precisely known value of the neutron lifetime will benefit both weak-interaction and Big Bang nucleosynthesis (BBN) theory. Combined with the measured neutron beta-decay asymmetry, the neutron lifetime yields the V_{ud} element of the Cabibbo–Kobayashi–Maskawa (CKM) quark mixing matrix. Non-unitarity of the CKM matrix would indicate physics beyond the Standard Model¹. In addition, the theoretical uncertainty in the predicted ⁴He abundance in the BBN model is dominated by the uncertainty in the neutron lifetime².

Many experiments, using different methods, have measured the neutron beta-decay lifetime^{5–9}. The most precise measurements use ultracold neutrons (UCN, neutrons with energies $E/k_b \leq 1$ mK, where k_b is Boltzmann's constant)¹⁰ confined in a 'bottle' made of materials with high UCN reflectivity. The largest uncertainties are due to UCN interactions with the material walls and fluctuations in the initial filling densities^{7,8}. Extrapolation methods are necessary to extract the neutron beta-decay lifetime. In another experiment, neutrons with specific trajectories and in a band of velocities (typically 10–14 m s⁻¹) were kept in a magnetic storage ring similar to those used in high-energy particle physics⁹. Since storage rings only confine neutrons in two dimensions, coupling between the large longitudinal momentum and small radial momentum by way of betatron oscillations allows neutrons to escape. In contrast, neutrons in magnetic traps are not susceptible to wall losses or betatron oscillations.

Static magnetic traps are formed by creating a magnetic field minimum in free space. The confining potential depth (D) of such a trap is determined by the magnetic moment of the trapped species (μ) and the difference (ΔB) between the magnitude of the field at the edge of the trap and at the minimum, $D = \mu\Delta B$. A particle in a low-field-seeking state (one with its magnetic moment anti-parallel with the local magnetic field vector) is pushed towards the trap minimum. Low-field-seeking particles with total energy less than D are energetically forbidden to leave the trapping region. For atoms and molecules with a magnetic moment of one Bohr magneton it is possible to produce trap depths of ~ 1 K. The trap depth for a neutron in the same trap would be only ~ 1 mK, because of its much smaller magnetic moment. Despite this difficulty, magnetic trapping of the neutron was proposed as early as 1961 by Vladimirov¹¹. His proposed technique was later used to confine neutrons using a combination of gravity and magnets¹². A separate effort to trap neutrons using a similar loading method to our work (but a different detection scheme) was unsuccessful because of the high temperature of the helium during the loading phase¹³.

Crucial to the utility of traps are the techniques used to load them. In order to catch a particle in a static conservative trap, its energy must be lowered while it is in the potential well. Atoms and molecules have been cooled and loaded into magnetic traps by scattering with either cryogenic surfaces^{14,15}, cold gases¹⁶ or photons from a laser beam (laser cooling)¹⁷. Neutrons, however, cannot be loaded by such methods because they cannot be excited optically and interact too weakly with atoms to be effectively cooled by a gas. Direct thermalization with a cold solid or liquid is generally precluded by the high probability for neutron absorption in the vast majority of materials.

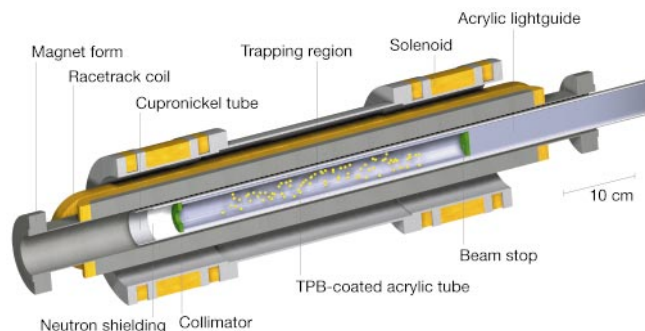


Figure 1 Half-section view of the neutron trapping apparatus. The trapping region is filled with isotopically pure ⁴He at a temperature ≤ 250 mK. The helium is contained within a cupronickel tube. A beam of cold neutrons passes through a series of teflon windows³⁰ and enters the helium from the left. It is collimated by a boron carbide ring, passes through the trapping region and is absorbed by a boron carbide beam stop. Approximately 1% of the 11 K neutrons scatter in the superfluid helium. Those neutrons (yellow) in the low-field-seeking spin state and with energy below the trap depth are magnetically confined. The rest of the scattered neutrons are absorbed by neutron-shielding material (boron

nitride) surrounding the trapping region. The magnetic trapping field is created by an assembly of superconducting magnets. Radial confinement is provided by a quadrupole constructed from four racetrack-shaped coils and axial confinement is provided by two sets of solenoids. Electrons from neutron beta-decay cause extreme-ultraviolet scintillations in the superfluid helium which are wavelength-shifted to the visible by a thin film of TPB-doped²³ polystyrene coated on the inside of an acrylic tube surrounding the trapping region. This tube is optically coupled to an acrylic light guide which transports the blue light to the end of the 250-mK region (to the right).

Our trapping of neutrons relies on a loading technique that employs the “superthermal process”¹⁸. A neutron with kinetic energy near 11 K (where the free-neutron and superfluid-helium dispersion curves cross) that passes through the helium-filled trapping region can lose nearly all of its energy through the creation of a single phonon. Neutrons that scatter to energies less than the trap depth (≤ 1 mK) and in the appropriate spin state are trapped. Upscattering of UCN in liquid helium by single-phonon absorption is highly suppressed due to the low density of 11 K phonons in the superfluid (proportional to $e^{-11\text{ K}/T}$). The rate of higher-order (multiple-phonon) upscattering processes is proportional to T^7 (ref. 19). The helium is maintained below a temperature of 250 mK to minimize upscattering of the trapped neutrons (total calculated upscattering rate per neutron $< 10^{-6} \text{ s}^{-1}$).

Another possible loss mechanism for trapped neutrons is nuclear absorption due to ^3He contamination of the superfluid ^4He bath. Unlike ^4He which does not absorb thermal neutrons at all, ^3He readily absorbs neutrons with a thermal cross section, $\sigma = 5,330 \times 10^{-24} \text{ cm}^2$. To minimize these losses, our superfluid-helium bath was isotopically purified ($^3\text{He}/^4\text{He} < 5 \times 10^{-13}$, ref. 20), giving a calculated absorption loss rate per neutron of $5 \times 10^{-6} \text{ s}^{-1}$. Thus, the trapped neutrons travel unimpeded through the helium and reside in the trap for a time limited primarily by the beta-decay lifetime of the neutron.

In our experiment, the superfluid helium is contained in a tube that is axially centred within a superconducting magnetic trap (see Fig. 1). The incident cold-neutron beam is collimated by a neutron-absorbing ring, passes through the trapping region, and is absorbed by the beam stop. As the beam traverses the trapping region, about 1% of the 11 K neutrons scatter in the helium. Some of these neutrons are trapped and the remainder are absorbed by shielding materials that surround the helium.

When a trapped neutron decays (into an electron, proton and anti-neutrino), the resulting high-energy electron (up to 782 keV) travels through the helium leaving a trail of ionized helium atoms. These ions quickly combine with neutral helium atoms to form excimer diatomic molecules, about half in singlet states and half in triplet states. About 15 singlet molecules are created per kilo-electronvolt of electron energy and the trail is 0–20 mm in length²¹. The singlet molecules decay within 10 ns, emitting a pulse of light in the extreme ultraviolet (EUV), $\lambda \approx 70\text{--}90 \text{ nm}$. This pulse of scintillation light is the signal of a neutron decaying in the trap. Radiative decay of the triplet states has a much longer lifetime, which we have measured to be $13 \pm 2 \text{ s}$ (ref. 22), and does not contribute to our signal.

The EUV scintillation light is wavelength-shifted to the visible and transported to the outside of the cryostat where it is detected. As shown in Fig. 1, the superfluid-helium-filled trapping region is surrounded by an acrylic tube. The inside surface of this tube is coated with a thin layer of polystyrene doped with the organic fluor tetraphenyl butadiene (TPB; ref. 23). The TPB converts the EUV scintillation light into blue light, a portion of which is internally reflected down the length of the tube²⁴. The tube is optically coupled to a solid light guide that transports the light to the end of the 250 mK region. The light then passes through a window at 4 K and into a second light guide which exits the Dewar. The light is split into two guides that are each coupled to a photomultiplier tube. Background from uncorrelated photons is suppressed by requiring coincident detection of at least two photoelectrons in each photomultiplier tube. This is required because of significant time-dependent luminescence (initial counting rate without coincident detection of $\sim 10^4 \text{ s}^{-1}$), induced by neutron absorption in the shielding materials. In a series of calibration runs (using a ^{113}Sn , 360-keV beta line source placed in the centre of the trapping region) we have determined that a total of 14.5 ± 1.5 photoelectrons per megaelectronvolt of beta energy are detected in a pulse 20 ns in width. We estimate that a coincidence signal is recorded for $31 \pm 4\%$ of

neutron decays in the trapping region.

In each experimental run, the cold-neutron beam is allowed to pass through the trapping region for 1,350 s, after which the beam is blocked and pulses of light are counted for 3,600 s. While the beam is on, neutrons interact with the helium and the number of UCN in the trap increases. After the beam is turned off, the photomultiplier tubes are turned on and a decreasing rate of scintillation events is observed. A background signal, consisting of both constant and time-varying components, obscures the trapped-neutron signal. These backgrounds are subtracted by taking data in two different kinds of run, trapping (signal + background) and non-trapping (background only), and subtracting the two. In a trapping run the magnetic field is on during the initial loading phase, so that UCN are confined in the trap, while in a non-trapping run the magnetic

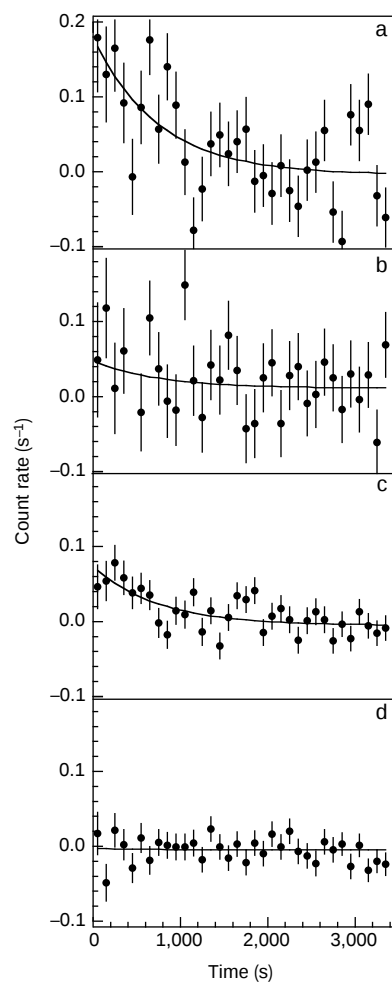


Figure 2 Counting rate as a function of time after the neutron beam is turned off (pooled background-subtracted data). The two sets of trapping data are fitted for five parameters: α_1 , α_2 , C_1 , C_2 and τ . The constant offset terms, C_1 and C_2 are found to be -0.004 ± 0.015 and -0.006 ± 0.006 respectively, both consistent with zero. The best-fit values for the initial counting rates for trapping data sets 1 and 2 are $\alpha_1 = 0.196 \pm 0.05 \text{ s}^{-1}$ and $\alpha_2 = 0.084 \pm 0.02 \text{ s}^{-1}$. The detection efficiency of $\epsilon = 31\% \pm 4\%$ combined with the known neutron lifetime, gives $N_1 = 560 \pm 160$ and $N_2 = 240 \pm 65$. The best fit value for the lifetime is $\tau = 750^{+330}_{-200} \text{ s}$. The results of this fit are shown on plots **a** and **c**. The ^3He data is fitted to four parameters: α_1 , α_2 , C_1 , C_2 and τ fixed at 750 s. Because no fitting parameter applies to both data sets, this is equivalent to two separate two-parameter fits, one to each data set. The amplitudes, $\alpha_1^{\text{He}} = 0.040 \pm 0.045$ and $\alpha_2^{\text{He}} = 0.005 \pm 0.016$, are both consistent with zero. The constant offset terms are $C_1 = 0.011 \pm 0.011$ and $C_2 = -0.006 \pm 0.004$. The ^3He data and fit are shown on plots **b** and **d**. **a**, Trapping data set 1 (23 pairs of runs at a trap depth of 0.76 mK). **b**, ^3He data set 1 (20 pairs at 0.76 mK). **c**, Trapping data set 2 (120 pairs at 0.50 mK). **d**, ^3He data set 2 (125 pairs at 0.50 mK).

field is initially off, so that no neutrons can be trapped. In a non-trapping run, the field is turned on during detection so that any residual luminescence (which depends on the magnetic field) will be properly subtracted. Equal numbers of each kind of run were taken, pooled and subtracted to give the background subtracted data.

Two sets of background subtracted data were collected: set 1 with a trap depth of 0.76 mK (Fig. 2a) and set 2 with a lower trap depth of 0.50 mK (Fig. 2c). (The lower trap depth was used because of problems with the magnet.) Most of the run-to-run variation in background rate is eliminated by excluding the first two pairs of runs in which the background rate is changing quickly due to activation of materials with lifetimes > 12 h. The remaining 23 pairs in set 1 (from about five days of running) and 120 pairs in set 2 (from about three weeks of running) are pooled and modelled as:

$$W_1 = \alpha_1 e^{-t/\tau} + C_1$$

$$W_2 = \alpha_2 e^{-t/\tau} + C_2$$

The subscripts 1 and 2 refer to sets 1 and 2, $\alpha_i = \epsilon N_i/\tau$, where N_i is the initial number of trapped neutrons and $i = 1$ or 2 , ϵ is the detection efficiency, and τ is the lifetime of neutrons in the trap. The constant C_i is present to account for the possible remaining effect of the changing background rate due to long-lived activation. However, in all of our fits the value of C_i is consistent with zero. The fit is performed simultaneously on the two data sets, minimizing the total χ^2 while varying five parameters: α_1 , α_2 , C_1 , C_2 and τ . The only parameter connecting the two data sets is τ . The best-fit values indicate $N_1 = 560 \pm 160$ and $N_2 = 240 \pm 65$. Calculations using the known beam flux, trap geometry and the theory of the superthermal process predict $N_1 = 480 \pm 100$ and $N_2 = 255 \pm 50$, in good agreement with the measured values. The best-fit value for the lifetime, $\tau = 750^{+330}_{-200}$ s, is consistent with the presently accepted value of the neutron beta-decay lifetime of 886.7 ± 1.9 s (ref. 5). All of the errors quoted correspond to a 68% confidence interval.

To verify that our signal is due to trapped neutrons, we doped the isotopically pure ^4He with ^3He at a concentration of 2×10^{-7} $^3\text{He}/^4\text{He}$. This amount of ^3He absorbs the trapped neutrons in less than 1 s without affecting anything else in the experiment (less than 1% of the 11 K neutrons are absorbed by ^3He). The ^3He background-subtracted data (see Fig. 2b and d) is consistent with zero, and differs from the signal data by more than four standard deviations. This confirms that our signal is due to trapped neutrons.

This work demonstrates the loading, trapping, and detection techniques necessary for performing a neutron-lifetime measurement using magnetically trapped UCN. Another important result is the direct confirmation of the theoretical prediction of the UCN density in the trap. Our value for the number of trapped neutrons at a trap depth of 0.76 mK corresponds to a density of 2 UCN cm^{-3} , compared to the density of 1 UCN cm^{-3} obtained in UCN material bottle experiments^{7,8} with a comparable UCN cut-off energy and higher flux reactor. Our measured density is consistent with previous measurements of the UCN production rate based on observation of upscattered UCN at higher temperatures²⁵ and confirms that the low number of UCN detected in ref. 26 was not due to an intrinsically low UCN production rate.

Magnetic trapping should allow significant improvement in the measurement of the neutron lifetime τ_n . There are several loss mechanisms specific to this method, including thermal upscattering and ^3He absorption. Both of these can be suppressed to a high level as discussed above. Another possible loss mechanism is neutron depolarization. When the neutron passes through a region where the field changes direction quickly compared to the neutron's spin precession frequency, the neutron can flip spin and be ejected from the trap. In our trap geometry, the trap has no zero-field regions and this loss rate can be minimized by the application of a bias field^{3,27}.

Using our method with a higher-flux source of cold neutrons, a

10^{-5} measurement of τ_n should be possible³. In the near term, improvements such as increased trap size and depth (resulting also in an improved ratio of signal-to-background events) should allow an improved accuracy measurement of τ_n at the 5×10^{-4} level. Other potential applications of superthermal UCN sources include searches for time-reversal violation²⁸ and new methods of neutron scattering especially suited for the study of large (biological-scale) molecules²⁹.

Received 5 August; accepted 16 November 1999.

- Dubbers, D. Test of the standard model with free neutron beta decay. *Nucl. Phys. A* **527**, 239–250 (1991).
- Lopez, R. E. & Turner, M. S. Precision prediction for the big-bang abundance of primordial ^4He . *Phys. Rev. D* **59**, 103502/1–14 (1999).
- Doyle, J. M. & Lamoreaux, S. K. On measuring the neutron beta-decay lifetime using ultracold neutrons produced and stored in a superfluid- ^4He -filled magnetic trap. *Europhys. Lett.* **26**, 253–258 (1994).
- Phillips, W. D. in *Laser Manipulation of Atoms and Ions* (eds Arimondo, E., Phillips, W. & Sturmia, F.) 289–344 (North Holland, Amsterdam, 1992).
- Caso, C. et al. Review of particle physics. *Eur. Phys. J. C* **3**, 1–794 (1998).
- Byrne, J. et al. A revised value for the neutron lifetime measured using a Penning trap. *Europhys. Lett.* **33**, 187–192 (1996).
- Mampe, W., Ageron, P., Bates, C., Pendlebury, J. M. & Steyerl, A. Neutron lifetime measured with stored ultracold neutrons. *Phys. Rev. Lett.* **63**, 593–596 (1989).
- Nesvizhevskii, V. V. et al. Measurement of the neutron lifetime in a gravitational trap and analysis of experimental errors. *Sov. Phys. JETP* **75**, 405–412 (1992).
- Paul, W., Anton, F., Paul, L., Paul, S. & Mampe, W. Measurement of the neutron lifetime in a magnetic storage ring. *Z. Phys. C* **45**, 25–30 (1989).
- Golub, R., Richardson, D. & Lamoreaux, S. K. *Ultra-Cold Neutrons* (Adam Hilger, Bristol, UK, 1991).
- Vladimirov, V. V. Magnetic mirrors, channels and bottles for cold neutrons. *Sov. Phys. JETP* **12**, 740–746 (1961).
- Abov, Yu. G., Borovl'ev, S. P., Vasil'ev, V. V., Vladimirov, V. V. & Mospan, E. N. Measurement of the time of storage of ultracold neutrons in a magnetic trap. *Sov. J. Nucl. Phys.* **38**, 70–73 (1983).
- Niehues, N. *Untersuchungen an einer magnetischen Flasche zur Speicherung von Neutronen. (Investigations on a magnetic bottle for the storage of neutrons.)* Thesis, Friedrich Wilhelm Univ. Bonn (1983).
- Hess, H. F. et al. Magnetic trapping of spin-polarized atomic hydrogen. *Phys. Rev. Lett.* **59**, 672–675 (1987).
- van Roijen, R., Berkhou, J. J., Jaakkola, S. & Walraven, J. T. M. Experiments with atomic hydrogen in a magnetic trapping field. *Phys. Rev. Lett.* **61**, 931–934 (1988).
- Doyle, J. M., Friedrich, B., Kim, J. & Patterson, D. Buffer-gas loading of atoms and molecules into a magnetic trap. *Phys. Rev. A* **52**, R2515–R2518 (1995).
- Chu, S. in *Laser Manipulation of Atoms and Ions* (eds Arimondo, E., Phillips, W. & Sturmia, F.) 239–288 (North Holland, Amsterdam, 1992).
- Golub, R. & Pendlebury, J. M. The interaction of ultra-cold neutrons (UCN) with liquid helium and a superthermal UCN source. *Phys. Lett. A* **62**, 337–339 (1977).
- Golub, R. et al. Operation of a superthermal ultracold neutron source and the storage of ultracold neutrons in superfluid helium-4. *Z. Phys. B* **51**, 187–193 (1983).
- Hendry, P. C. & McClintock, P. V. E. Continuous flow apparatus for preparing isotopically pure ^4He . *Cryogenics* **27**, 131–138 (1987).
- Adams, J. S., Kim, Y. H., Lanou, R. E., Maris, H. J. & Seidel, G. M. Scintillation and quantum evaporation generated by single monoenergetic electrons stopped in superfluid helium. *J. Low Temp. Phys.* **13**, 1121–1128 (1998).
- McKinsey, D. N. et al. Radiative decay of the metastable He_2 ($a_3 \Sigma_u^+$) molecule in liquid helium. *Phys. Rev. A* **59**, 200–204 (1999).
- McKinsey, D. N. et al. Fluorescence efficiencies of thin scintillating films in the extreme ultraviolet spectral region. *Nucl. Instrum. Methods B* **132**, 351–358 (1997).
- Habicht, K. *Scintillationen in flüssigem Helium—ein Detektor für ultrakalte Neutronen. (Scintillations in liquid helium—a detector for ultracold neutrons.)* Thesis, Tech. Univ. Berlin (1998).
- Kilvington, A. I., Golub, R., Mampe, W. & Ageron, P. Scattering of ultra-cold neutrons (UCN) by superfluid helium at temperatures around 1 K. *Phys. Lett. A* **125**, 416–420 (1987).
- Yoshiki, H. et al. Observation of ultracold-neutron production by 9-Å cold neutrons in superfluid helium. *Phys. Rev. Lett.* **68**, 1323–1326 (1992).
- Happer, W. Optical pumping. *Rev. Mod. Phys.* **44**, 169–249 (1972).
- Golub, R. & Lamoreaux, S. K. Neutron electric-dipole moment, ultracold neutrons and polarized ^3He . *Phys. Rep.* **237**, 1–62 (1994).
- Golub, R. Ultracold neutrons: their role in studies of condensed matter. *Rev. Mod. Phys.* **68**, 329–347 (1996).
- Butterworth, J. S. et al. A removable cryogenic window for transmission of light and neutrons. *Rev. Sci. Instrum.* **69**, 3998–3999 (1998).

Acknowledgements

We thank J. M. Rowe, D. M. Gilliam, G. L. Jones, J. S. Nico, N. Clarkson, G. P. Lamaze, C. Chin, C. Davis, D. Barkin, A. Black, V. Dinu, J. Higbie, H. Park, R. Ramakrishnan, I. Siddiqi and G. Brandenburg for their help with this project. We thank P. McClintock, D. Meredith and P. Hendry for supplying the isotopically pure helium. We acknowledge the support of the NIST, US DOC, in providing the neutron facilities used in this work. This work is supported in part by the US NSF. The NIST authors acknowledge the support of the US DOE.

Correspondence and requests for materials should be addressed to P.R.H. (e-mail: paul.huffman@nist.gov).

Reverse microemulsion synthesis of nanostructured complex oxides for catalytic combustion

Andrey J. Zarur & Jackie Y. Ying

Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307, USA

Catalysts play an important role in many industrial processes, but their use in high-temperature applications—such as energy generation through natural gas combustion, steam reforming and the partial oxidation of hydrocarbons to produce feedstock chemicals—is problematic. The need for catalytic materials that remain stable and active over long periods at high operation temperatures, often in the presence of deactivating or even poisoning compounds, presents a challenge. For example, catalytic methane combustion, which generates power with reduced greenhouse-gas and nitrogen-oxide emissions^{1–3}, is limited by the availability of catalysts that are sufficiently active at low temperatures for start-up and are then able to sustain activity and mechanical integrity at flame temperatures as high as 1,300 °C. Here we use sol–gel processing in reverse microemulsions to produce discrete barium hexaaluminate nanoparticles that display excellent methane combustion activity, owing to their high surface area, high thermal stability and the ultrahigh dispersion of cerium oxide on their surfaces. Our synthesis method provides a general route to the production of a wide range of thermally stable nanostructured composite materials with large surface-to-volume ratios^{4–6} and an ultrahigh component dispersion that gives rise to synergistic chemical and electronic

effects^{7,8}, thus paving the way to the development of catalysts suitable for high-temperature industrial applications.

Combustion catalysts can be gainfully employed to stabilize 'lean' flames (with low fuel to air ratios) at temperatures significantly lower than those required in non-catalytic combustion, thereby minimizing the production and emission of nitrogen oxide (NO_x) species^{1–3} that contribute to acid rain and smog formation. A viable catalyst system needs to be active at low temperatures for start-up and transient periods, and to sustain activity and mechanical integrity at flame temperatures as high as 1,300 °C. In the case of methane, light-off (defined as 10% conversion of the fuel stream) should ideally be achieved at temperatures of about 400 °C.

Catalyst systems based on noble metal oxides, such as PdO, have been studied for catalytic combustion applications⁹, but they generally suffer from deactivation at 700–800 °C owing to loss of active species and sintering. Catalysts that use oxides of non-noble metals^{10–12}, especially in the form of complex oxides such as aluminates and perovskites, seem more promising, with systems based on barium hexaaluminate (BHA) having received particular attention¹³. BHA, which has a complex crystal structure giving rise to highly anisotropic crystal growth that is usually suppressed along the *c*-axis¹⁴, has been synthesized by sol–gel processing. However, this material has to be heated to 1,300 °C to achieve crystallization, retains a surface area of only 15 m² g^{–1} after calcination¹⁵ and requires about 700 °C for light-off of a stream of 1-vol% methane (CH₄) in air¹⁶.

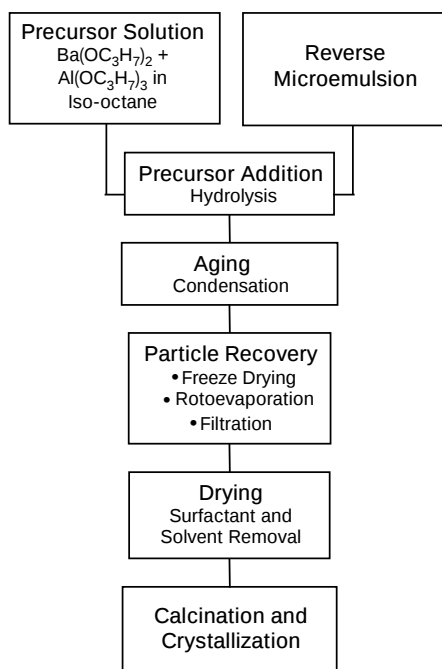


Figure 1 Diagram of the reverse-microemulsion-mediated sol–gel processing of barium hexaaluminate (BHA) nanoparticles. The barium and aluminium alkoxide precursors were dissolved in iso-octane. A reverse microemulsion was used as hydrolysing medium for the alkoxide precursors. The hydrolysed materials were aged within the reverse micelles to promote condensation. The nanoparticles were recovered from the reverse microemulsion preferentially via freeze drying. The surfactants and solvents were removed by supercritical drying and calcination. The materials were then heat treated at high temperatures to promote crystallization.

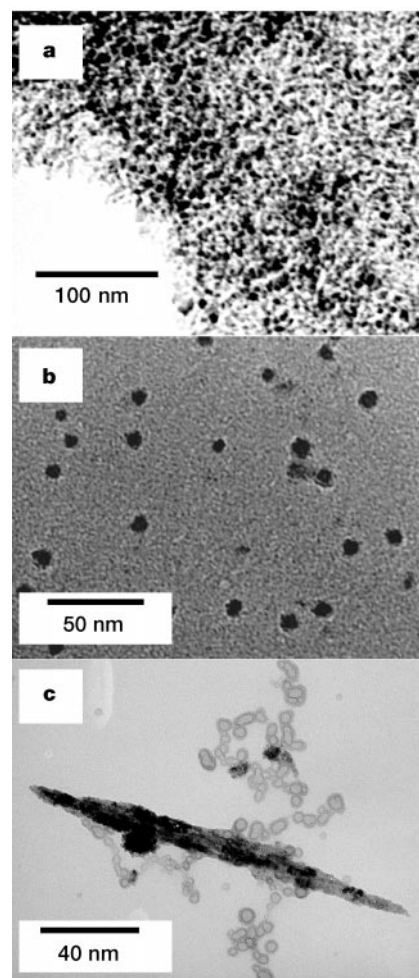


Figure 2 Characterization of uncalcined BHA materials. Transmission electron micrographs: **a**, gel-like BHA recovered from a reverse microemulsion with 1-wt% H₂O; **b**, discrete BHA nanoparticles formed in a reverse microemulsion with 15-wt% H₂O; **c**, a mixture of BHA filament-like and spherical particles synthesized in a reverse microemulsion with 30-wt% H₂O.

BHA is difficult to synthesize in a nanocrystalline form, due in part to the different reactivity of the barium and aluminium precursors. The compositional heterogeneity in the as-synthesized material requires crystallization via solid-state reaction at very high temperatures, which induces loss of surface area associated with grain growth. To achieve BHA nanocrystals at relatively low temperatures, we developed a reverse-microemulsion-mediated synthesis (see Fig. 1 and Methods). In this technique, nanometre-sized aqueous micelles dispersed in an oil phase are used as nanoreactors for controlled hydrolysis and condensation of Ba and Al alkoxides. Unlike conventional sol-gel processing, the reaction rate in this approach is controlled by the diffusion of precursors from the oil phase to the aqueous domains, instead of the hydrolysis of one of the precursors. Despite the different hydrolysis rates of Ba and Al alkoxides, chemical homogeneity can be attained with the mediation of the reverse microemulsion since the alkoxides have similar diffusivities in the oil phase.

To tailor the particle morphology, structure, and surface area of BHA nanoparticles, synthesis parameters such as the composition of the reverse microemulsion, water to alkoxide ratio, aging time, powder recovery and drying techniques were carefully controlled. The composition of the reverse microemulsion governs the shape and size of the reverse micellar domains, as well as the phase stability. This is reflected by the morphology of the powders

obtained from the reverse-microemulsion-mediated synthesis. Reverse microemulsions with a water content of <5 wt% underwent phase separation shortly after the addition of the precursor solution. The powders produced could not be stably suspended in the resulting mixture, and were agglomerated to form a gel-like structure (Fig. 2a). The ultrafine pores of this material were easily collapsed during calcination, yielding coarse-grained materials with surface areas <20 m² g⁻¹, similar to those obtained from conventional sol-gel processing.

With an intermediate water content of 5–20 wt%, discrete spherical particles of 3–10 nm were obtained from the reverse-microemulsion-mediated synthesis (Fig. 2b). These nanoparticles could be crystallized directly to the desired BHA phase at a relatively low temperature of 1,050 °C, and sustained surface areas in excess of 100 m² g⁻¹ after calcination to 1,300 °C. The excellent thermal stability of these nanoparticles might be attributed to the successful separation of the crystallization and grain-growth processes. Conventional materials undergo BHA crystallization between 1,250 and 1,350 °C where a strong driving force for grain growth exists. Reverse-microemulsion-derived BHA nanoparticles possess superior chemical homogeneity, so that crystallization could be achieved at lower temperatures. Upon crystallization, grain growth at higher temperatures was suppressed, owing to the anisotropic crystal growth of BHA.

For systems with >20 wt% water, significant percolation of the micellar domains might occur, leading to a mixture of liquid-crystalline phases. The materials recovered from these systems showed spherical and elongated particle morphologies (Fig. 2c). These could be crystallized at 1,100–1,150 °C, and possessed a moderately high surface area of 40–60 m² g⁻¹ after calcination at 1,300 °C.

The water to alkoxide ratio is an important parameter for controlling the relative rates of hydrolysis and condensation reactions¹⁷. In general, higher levels of excess water lead to smaller particle size and increased surface area after calcination. At a water to alkoxide ratio of 1–10 times the stoichiometric value, particle growth was significant, as reflected by the change in appearance of the reverse microemulsion from transparent to opaque white. After a short aging period, the medium underwent phase separation, resulting in highly agglomerated materials. At a water to alkoxide ratio ≥ 100 times the stoichiometric value, the reverse microemulsion remained transparent or translucent throughout the aging process; uniform, discrete particles were recovered upon drying. After calcination to 1,300 °C, these nanoparticles retained surface areas of 40–160 m² g⁻¹, depending on the reverse-microemulsion composition and the recovery technique used.

Particle growth may be minimized by short aging periods. However, appropriate aging was essential to ensure completion of the condensation reaction. Condensation of the surface hydroxyl groups led to particles that were less susceptible to agglomeration in subsequent powder processing. We determined that the optimal aging period for BHA systems was 24–48 h.

Conventional techniques (such as filtration) were ineffective at recovering the materials prepared by reverse-microemulsion-mediated processing. Freeze drying was found to allow successful recovery of BHA nanoparticles with high yields while preserving their discrete, nanometre-sized particle morphology. To remove the residual surfactants and volatiles in the recovered particles, samples were subjected to drying in an oven or under supercritical conditions. Supercritical drying prevented particle agglomeration and growth during the removal of organics. Materials recovered by freeze drying and purified by supercritical drying retained surface areas as high as 160 m² g⁻¹ after calcination at 1,300 °C. Crystallization was achieved by 1,050 °C, and grain growth was minimal after further heat treatment. The final BHA particle size was only 30 nm, as shown in Fig. 3a.

Our nanocrystalline BHA materials showed excellent catalytic combustion activity. Light-off of a stream of 1-vol.% CH₄ in air at a

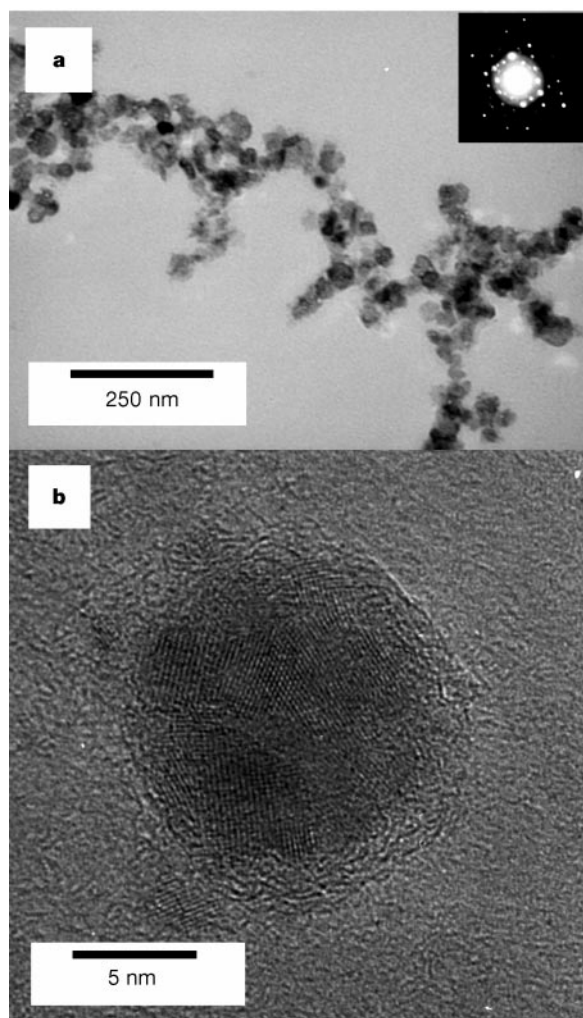


Figure 3 Characterization of calcined BHA nanoparticles. **a**, Transmission electron micrograph and electron diffraction pattern (inset) of reverse-microemulsion-derived BHA nanoparticles after calcination at 1,300 °C. **b**, High-resolution transmission electron micrograph of BHA nanoparticles with 10-wt% nanocrystalline CeO₂ surface coating after calcination at 800 °C.

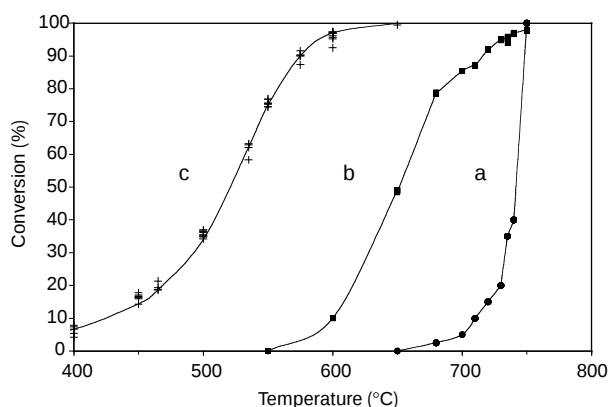


Figure 4 Catalytic methane combustion experiments. Catalytic methane oxidation activity: **a**, sol-gel-derived conventional BHA; **b**, reverse-microemulsion-derived BHA nanoparticles; **c**, reverse-microemulsion-derived CeO₂-BHA nanocomposite. The reactant stream consisted of 1-vol.% CH₄ in air (space velocity, 60,000 h⁻¹).

high space velocity of 60,000 h⁻¹ was observed at 590°C for nanocrystalline BHA (Fig. 4b), compared to 710°C required for conventional BHA (Fig. 4a). The nanocrystalline BHA catalyst provided full CH₄ conversion over a wide temperature range of 710–1,300°C, with no signs of deactivation at the elevated reaction temperatures.

Introduction of surface-active deposits onto the BHA nanoparticles further increased their catalytic activity. As ceria does not react with baria or alumina under our synthesis conditions, we were able to deposit up to 25-wt% ceria onto BHA nanoparticles without forming a separate ceria domain. The ceria deposits on BHA preserved a nanocrystalline morphology even after calcination at 1,300°C, whereas unsupported ceria would have undergone severe sintering and grain growth beyond 100 nm by 700°C. High-resolution transmission electron microscopy (HR-TEM) illustrated that ceria (CeO₂) nanocrystals were achieved with ultrahigh dispersion on the surface of BHA nanoparticles using our synthesis approach (Fig. 3b). The ceria grain size (~6 nm) observed in HR-TEM is in good agreement to that calculated from peak-broadening analysis of the X-ray diffraction (XRD) pattern. With the use of the stable BHA nanoparticles as support materials, a fine grain size of 18 nm was preserved for ceria after extended calcination at 1,100°C in the presence of 15-vol.% H₂O.

The CeO₂-BHA nanoparticles allowed light-off of a stream of 1-vol.% CH₄ in air at a low temperature of ~400°C (Fig. 4c), compared to 540°C and 690°C required for Mn- and Co-substituted BHA systems¹⁶. The apparent activation energy of the CeO₂-BHA nanoparticles for combustion of <4-vol.% CH₄ in air was found to be ~145 kJ mol⁻¹. In comparison, the conventional 0.5-wt% PdO/Al₂O₃ had an activation energy only slightly lower (116 kJ mol⁻¹). The CeO₂-BHA catalysts attained full CH₄ conversion by 600°C, and no hysteresis was observed in the catalyst performance during heating-cooling-restarting cycles. We believe that the excellent low-temperature catalytic activity of this system arises in part from the ultrahigh CeO₂ dispersion on BHA nanoparticles. CeO₂-based materials have not been examined widely as combustion catalysts, since other metal-oxide systems (such as Co₂O₃ and Mn₃O₄) possess higher specific activities¹⁸. However, the use of the transition-metal-oxide catalysts has been limited to ~1,000°C, above which severe grain growth and sintering occur. In contrast, our CeO₂-BHA system not only achieved light-off at low temperatures, but also preserved full methane conversion activity at temperatures in excess of 1,100°C and in the presence of H₂S and water vapour. This combination of low-temperature catalytic activity, high-temperature thermal stability, and poisoning resistance render our catalysts interesting for potential

practical applications in ultralean catalytic combustion of methane. Provided the composite oxides do not react with each other under our synthesis conditions, the nanocomposite design and the ultrahigh component dispersion achieved by our reverse-microemulsion-mediated synthesis can be broadly applied to the development of other complex-oxide catalysts with high surface reactivity and thermal stability for a variety of catalytic applications. □

Methods

The reverse microemulsions of water in oil (for example, iso-octane) were prepared with a mixture of surfactants, such as polyethylene oxide adducts and linear alcohols. Typical microemulsion compositions consist of 1–40 wt% water, 5–30 wt% surfactants, and 60–90 wt% hydrocarbons. Ba(OC₃H₇)₂ and Al(OC₃H₇)₃ were dissolved in iso-octane to form the precursor solution, which was then slowly added (5–10 ml min⁻¹) to the reverse microemulsion at room temperature. The water to alkoxide ratios were varied from 1 to 1,000 times the stoichiometric value (corresponding to 19 moles of water per mole of BaO·6Al₂O₃ derived). The hydrolysed mixture was then allowed to age for 1–72 h.

The BHA nanoparticles could be modified through the introduction of surface deposits of rare-earth and transition-metal oxides. Nitrate or acetate salts of the desired cations were introduced to the BHA nanoparticles during their aging in the reverse microemulsion medium 12–24 h after hydrolysis had taken place. This approach allowed the deposition of up to 25 wt% of various metal oxides, including those of cerium, cobalt, manganese, and lanthanum.

The materials generated in the reverse microemulsion could be recovered either by filtration, freeze drying, or phase separation (by cooling of the reverse microemulsion). The nanometre-sized particles were difficult to collect with high yields by filtration or centrifugation. Roto-evaporation resulted in heavy particle agglomeration due to the high temperatures needed to remove the surfactants. In contrast, freeze drying consisted of spraying the reverse microemulsion into a Dewar containing liquid nitrogen. The frozen system was then placed under vacuum, and heated slowly to remove iso-octane, water, and the lighter fractions of surfactants by sublimation. After recovery, the powders were subjected to oven or supercritical drying to remove the residual surfactants and volatiles. The materials were then calcined to 800°C or 1,300°C for the modified and pure BHA nanoparticles, respectively, before catalytic studies. These materials were shown by elemental analysis to be free of surfactants and volatiles (with <0.1 wt% carbon content).

Catalytic studies were conducted in a flow reactor under isothermal conditions. Methane and air flows were controlled by independent mass flow controllers (MKS) to vary the CH₄ to O₂ ratio. The effluent stream was analysed using a Hewlett-Packard gas chromatograph with a mass selective detector.

Received 4 May; accepted 21 October 1999.

- Pfefferle, L. D. & Pfefferle, W. C. Catalysis in combustion. *Catal. Rev. Sci. Eng.* **29**, 219–267 (1987).
- Linan, A. & Williams, F. A. *Fundamental Aspects of Combustion* (Oxford, New York, 1993).
- Lunsford, J. H. The role of surface generated gas-phase radicals in catalysis. *Langmuir* **5**, 12–16 (1989).
- Karch, J., Birringer, R. & Gleiter, H. Ceramics ductile at low temperature. *Nature* **330**, 556–558 (1987).
- Gleiter, H. Nanocrystalline materials. *Prog. Mater. Sci.* **33**, 223–315 (1989).
- Siegel, R. W. Cluster-assembled nanophase materials. *Annu. Rev. Mater. Sci.* **21**, 559–578 (1991).
- Ying, J. Y. & Tschöpe, A. Synthesis and characterization of non-stoichiometric nanocrystalline cerium oxide-based catalysts. *Chem. Eng. J.* **64**, 225–237 (1996).
- Ying, J. Y., Tschöpe, A. & Levin, D. Synergistic effects and catalytic properties tailored by nanostructure processing. *Nanostruct. Mater.* **6**, 237–246 (1995).
- McCarty, J. G. Kinetics of PdO combustion catalysis. *Catal. Today* **26**, 283–293 (1995).
- Trimm, D. L. Materials selection and design of high temperature catalytic combustion units. *Catal. Today* **26**, 231–238 (1995).
- Berg, M. & Järäs, S. G. Catalytic combustion of methane over magnesium oxide. *Appl. Catal. A* **114**, 227–241 (1994).
- Matsuura, I., Hashimoto, Y., Takayasu, O., Nitta, K. & Yoshida, Y. Heat-stable ultrafine single-crystal magnesium oxide and its character as a support material for high-temperature combustion catalysts. *Appl. Catal. A* **74**, 273–280 (1991).
- Groppi, G., Bellotti, M., Cristiani, C., Forzatti, P. & Villa, P. L. Preparation and characterization of hexaaluminate-based materials for catalytic combustion. *Appl. Catal. A* **104**, 101–108 (1993).
- Park, J.-G. & Cormack, A. N. Crystal/defect structures and phase stability in Ba hexaaluminates. *J. Solid State Chem.* **121**, 278–290 (1996).
- Machida, M., Eguchi, K. & Arai, H. Preparation and characterization of large surface area BHA. *Bull. Chem. Soc. Jpn* **61**, 3659–3665 (1988).
- Machida, M., Eguchi, K. & Arai, H. Catalytic properties of BaAl₁₁O_{19-α} for high-temperature catalytic combustion. *J. Catal.* **120**, 377–386 (1989).
- Scherer, G. W. & Brinker, C. J. *Sol-Gel Science: The Physics and Chemistry of Sol-Gel Processing* (Academic, Boston, 1990).
- Zwinkels, M. F. M., Järäs, S. G., Menon, P. G. & Griffin, T. A. Catalytic materials for high-temperature combustion. *Catal. Rev. Sci. Eng.* **25**, 319–358 (1993).

Acknowledgements

We thank H. H. Hwu for technical assistance and M. Frongillo for assistance with the high-resolution electron microscopy studies. This work was supported by the David and Lucile Packard Foundation.

Correspondence and requests for materials should be addressed to J.Y.Y. (e-mail: jyying@mit.edu).

Forecasting Andean rainfall and crop yield from the influence of El Niño on Pleiades visibility

Benjamin S. Orlove^{*†}, John C. H. Chiang[†] & Mark A. Cane[†]

^{*} Department of Environmental Science and Policy, University of California, Davis, California 95616, USA

[†] Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York 10964, USA

Farmers in drought-prone regions of Andean South America have historically made observations of changes in the apparent brightness of stars in the Pleiades around the time of the southern winter solstice in order to forecast interannual variations in summer rainfall and in autumn harvests. They moderate the effect of reduced rainfall by adjusting the planting dates of potatoes, their most important crop¹. Here we use data on cloud cover and water vapour from satellite imagery, agronomic data from the Andean altiplano and an index of El Niño variability to analyse this forecasting method. We find that poor visibility of the Pleiades in June—caused by an increase in subvisual high cirrus clouds—is indicative of an El Niño year, which is usually linked to reduced rainfall during the growing season several months later. Our results suggest that this centuries-old method² of seasonal rainfall forecasting may be based on a simple indicator of El Niño variability.

We reviewed anthropological accounts of indigenous Aymara- and Quechua-speaking farmers of the Peruvian and Bolivian Andes (hereafter central Andes). In 12 villages^{3–16} (see Fig. 1a, Table 1), the inhabitants observe the Pleiades in late June in order to forecast the weather during the growing season (October–May). Observations often begin around 15 June and culminate on 24 June, the festival of San Juan. They occur an hour or two before dawn, when the Pleiades are located low over the horizon to the northeast.

Four different attributes of the Pleiades were reported (Table 1). Two attributes are directly related to the relative clarity or transparency of the sky: the brightness of the cluster and the timing of the heliacal rise (the date of the first appearance in the eastern pre-dawn sky). The third attribute, the apparent size of the Pleiades, may also be related to atmospheric clarity. On nights when the dimmest stars are visible, the Pleiades will appear to be 25% larger in diameter than on less clear nights (Fig. 2). The fourth attribute, the relative position of the brightest star in the Pleiades, is reported in various ways in different villages. Our interpretation of this attribute is

uncertain, but the brightest star in the cluster may appear to shift its position when the dimmest stars are no longer visible.

The villagers use these attributes to forecast the timing and quantity of rains and to estimate the size of the harvest, concentrated between March and May of the following year. The attributes associated with clearer skies indicate earlier and more abundant rains and larger harvests, while the opposite is linked to less clear skies. If poor rains are predicted, villagers postpone the planting of potatoes. A shallow-rooted crop, potatoes are most vulnerable to drought at planting, when low soil moisture inhibits root formation¹⁷, and again a few weeks later when the sprouts have depleted the residual moisture in the seed tubers. Lack of soil moisture at that time reduces the number of stems per plant and the overall yield¹⁸. By delaying for 4–6 weeks after the usual October–November planting time, the farmers reduce these risks by starting the potato crop in months of higher rainfall.

This forecasting system is likely to be more than four centuries old. The Incas, who unified the central Andes in the fifteenth century and were in turn conquered by the Spaniards in 1532, worshipped and closely observed the Pleiades¹⁹. It is not established that they used its appearance to forecast weather, but several sources do document such forecasts soon after the conquest. In the late sixteenth century, some central Andean villages celebrated the Pleiades in June²⁰, some noted the date of its heliacal rise²¹, and some made forecasts that linked a large apparent size of the Pleiades with good harvests and a small size with poor harvests². The general importance of the Pleiades in Andean astronomy is suggested by its prominent position in a seventeenth-century cosmological chart²² (see Supplementary Information).

We now examine whether the visibility of the Pleiades from the central Andes in June is correlated with harvests in the following year. We use high cloud as a measure of atmospheric transparency (see below). The International Satellite Cloud Climatology Project (ISCCP) nadir-viewing satellite cloud data set provides data on high cloud cover over the central Andes²³. For a measure of harvests, we use yield for Puno department, located near the centre of the region in which forecasts are made. Figure 1a shows, as contours, the rank correlation between high cloud and potato harvest, indicative of the degree of correspondence between these two variables. Because of the low angle of the Pleiades at the time of observation, the relevant region is about 175 km to the northeast of the villages, where correlations are in the –0.5 to –0.6 range. In the five cases for which independent data on forecast and outcome are available, the forecasts have been accurate (see Supplementary Information).

To account for this correlation, we examine the links of high cloud amount and potato harvest, as well as precipitation, to El Niño/Southern Oscillation (ENSO). ENSO is an interannual climate fluctuation originating from large-scale dynamical interactions

Table 1 Village data on Pleiades observations

Village	Department	Country	Latitude (°S)	Longitude (°W)	Elevation (m)	Date			Attribute of Pleiades				Modification of Planting dates?	Ref.
						Early June observation	24 June observation	After 24 June	Brightness	Date of first appearance	Size	Position of brightest star		
Concepción	Junín	Peru	11.9	75.2	3,300	Not mentioned	Yes	No	No	Not mentioned	No	Yes	Yes	3
Misminay	Cusco	Peru	13.0	72.0	3,400	No	Yes	No	Yes	No	Yes	Yes	Yes	4
Quispillacta	Ayacucho	Peru	13.6	74.4	3,300	No	Yes	No	Yes	No	Yes	Yes	Yes	5
Kauri	Cusco	Peru	13.6	71.5	4,300	No	Yes	Yes	No	No	No	Yes	Yes	6
Sallaq	Cusco	Peru	13.7	71.6	3,200	Not mentioned	Yes	No	No	Not mentioned	No	Yes	Yes	7
Sicuaní	Cusco	Peru	14.3	71.2	3,550	Not mentioned	Yes	No	Yes	Not mentioned	Yes	No	Not mentioned	5, 8
Marangani	Cusco	Peru	14.4	71.1	3,600	No	No	Yes	Yes	No	Yes	No	Not mentioned	9
Cuyo Cuyo	Puno	Peru	14.5	69.5	3,750	Yes	Yes	No	Yes	Yes	No	No	Yes	10, 11
Qorpa	La Paz	Bolivia	16.5	68.7	3,850	Yes	Yes	No	Yes	Yes	No	Yes	Yes	12
Irpa Chico	La Paz	Bolivia	16.7	68.4	3,800	Yes	No	No	No	Yes	No	No	Yes	13, 14
Chambi Grande	La Paz	Bolivia	17.6	68.4	3,700	Not mentioned	Yes	No	Yes	Not mentioned	Yes	No	Yes	15
Chayantaka	Potosí	Bolivia	18.5	66.5	3,800	Yes	Yes	No	Yes	Yes	Yes	No	Yes	16

Data collected from the 12 villages, as reported in the anthropological accounts of farmers in the Peruvian and Bolivian Andes^{3–16}. The villages are listed in the order that corresponds to their numbering in Fig. 1a. Ten of the 12 villages have a Pleiades observation on 24 June. The most common attribute of the Pleiades reported is the brightness (8 villages). The size of the cluster and the position of the brightest star are observed in half of the villages (6 villages each), and is the date of first appearance (4 of the 8 villages for which data are reported). A large majority (at least 10) of the villages modifies their planting dates in response to the forecast.

between the ocean and atmosphere of the tropical Pacific, and has the largest effect on short-term global climate variability of any such fluctuation. ENSO influences a number of atmospheric variables in the central Andes. Atmospheric extinction data measured by the Stratospheric Aerosol and Gas Experiment II (SAGE II) solar occultation instrument show that tropic cirrus cloud cover increases in warm ENSO years relative to cold ENSO years over the central Andes²⁴. At the representative altitude of 14.5 km, the cloud amount increased from around 35% in cold ENSO years to 50% in warm ENSO years. Though too much high cloud could totally mask out the Pleiades, the SAGE II data show that most tropical high cloud is subvisual (with optical thickness <0.03) from just over 50% of cloud amount at the 10.5-km level to almost 100% above 16 km. We estimate the average optical thickness of the 10–19 km cloud layer to be between 0.05 and 0.35, with the major uncertainty due to uncertainty in the extinction coefficients of the optically thicker clouds in this layer. Change in cloud amount between cold and warm ENSO years thus leads to an estimated optical thickness change between 0.01 and 0.1, yielding a change in Pleiades brightness of around 0.1 to 1 magnitude. This shift is sufficient to visibly reduce the apparent brightness of the Pleiades between cold and warm years.

Increase in cloud frequency in the SAGE II data set is corroborated by high-cloud data from the ISCCP dataset²³ (Fig. 1b). Optically thin (<0.1 at $0.6 \mu\text{m}$) clouds are at the detection limit

for the ISCCP²⁵, but even if the optically thin clouds are not detected, the high-cloud correlation with ENSO implies more moist conditions during warm years and an increased likelihood of thin clouds.

We considered other factors that might influence atmospheric visibility in the central Andes in June. Using a reanalysis data set²⁶, we found that upper-level winds increase in warm ENSO years. Though the accompanying decrease in vertical stability might lead to more turbulence, this change would have little direct effect on visibility. ENSO is also significantly correlated with total column-integrated precipitable water northeast of the villages, as estimated by mean monthly data from the Total Ozone Vertical Sounder (TOVS)²³. The increase in total precipitable water from cold to warm ENSO years is around 10–20% of the mean amount, and may contribute to the formation of thin high cloud. The reanalysis data set²⁶ indicates that the surface relative humidity of about 60–70% increases by about 5–10% in warm years. Standard formulas²⁷ indicate that this increase would alter apparent magnitude by about 0.1, a level too small to be appreciated by the naked-eye observer. We conclude that the significant changes in visibility are attributable to changes in high cloud alone. However, our confidence in the link between ENSO and shifts in cloud cover is bolstered by the consistent relationship found within independent data sets between ENSO and other atmospheric variables.

We note that synoptic variability in atmospheric conditions

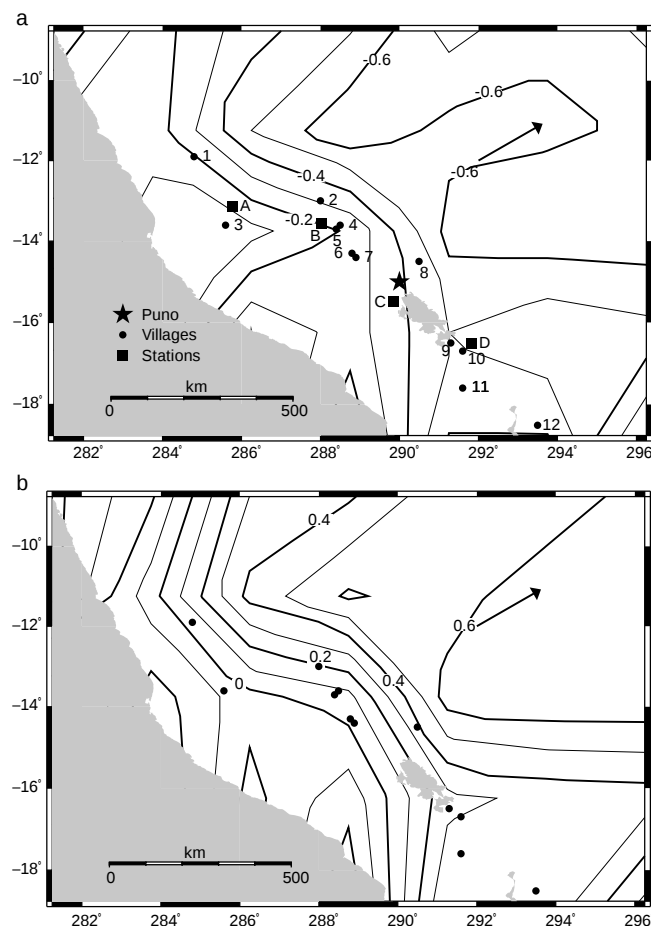


Figure 1 Correlations with central Andean high cloud in late June. Contours show rank correlations with the ISCCP C29 GMT anomalous high-cloud amount over late June (see text). Filled circles indicate location of the 12 villages listed in Table 1, filled squares the location of the four meteorological stations, and the star the approximate geographical centre of Puno department. The arrow indicates the azimuth of Pleiades observations during late June. Late June anomalous high cloud is computed by linearly interpolating the June and July anomalous monthly cloud data to 24 June, assuming that monthly

mean data are centred on the 15th of the month. The exceptions are 1983, where we take July anomalies only, and 1991, where we take June anomalies only (ISCCP data runs from July 1983 to June 1991). **a**, Rank correlation with potato yield in Puno department in the year following the June cloud data. Annual potato yield is computed from August to the following July. **b**, Rank correlation with the NINO3 index (the sea surface temperature anomaly averaged over 5° S–5° N, and 150° W–90° W) averaged over November to January (NINO3NDJ) following June cloud data.

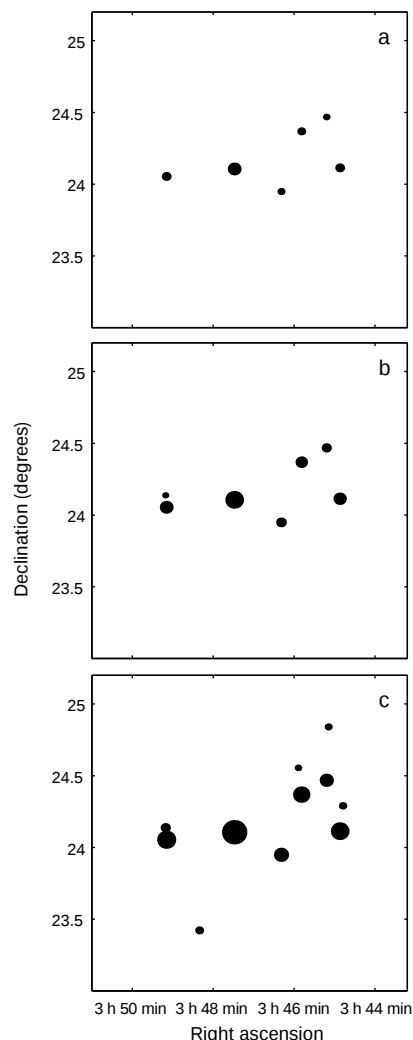


Figure 2 A diagram of the stars in the Pleiades of visual magnitude 6.0 or brighter. Although human visual acuity is often reduced at high elevations above sea level, suggesting that a lower threshold would be appropriate in this case, we note that hypoxia, the cause of this diminution, is not a problem for the populations under study³⁰. In this diagram, the size of each star is proportional to its intensity. We show three frames representing the Pleiades in steps of a 0.75 change in apparent magnitude, to show the effect of different observing conditions. The stars can be grouped into three brightness classes. The brightest group consists of only one star (Alcyone or 25 Tauri) of magnitude 2.90. There is a gap of 0.72 magnitude between this star and the intermediate group, of magnitude 3.62–4.30. A gap of 0.79 magnitude separates this group from the dimmest group, of magnitude 5.09–5.80. The axis of the six brightest stars in the Pleiades has an apparent breadth of 59 arcmin. For comparison, the moon's apparent diameter is less than 32 arcmin.

implies that a single viewing on 24 June at the festival of San Juan could be misleading. It is a common but not universal practice (Table 1) to make observations over a number of days, increasing the likelihood of registering the ENSO signal correctly. Additional research with a larger sample of villages would be needed to ascertain the range of observing practices. Are they near 'optimal' for the purpose of prediction, or are they strongly shaped by social and cultural factors?

That drought conditions in the central Andes are associated with El Niño has been documented^{28,29}. To further analyse this relation, we examined precipitation data between July 1962 and June 1988 from the 4 stations with nearly complete records representative of the region occupied by the 12 villages: Ayacucho, Cusco and Juliaca in Peru, and La Paz in Bolivia (Fig. 1a). After removing the

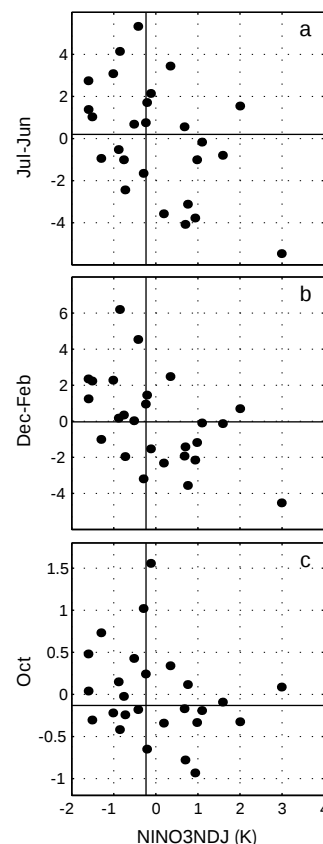


Figure 3 Central Andean rainfall anomalies and El Niño. Shown are scatterplots of a central Andes precipitation anomalies index versus the NINO3NDJ index; the latter is the NINO3 index averaged over November–January. All timeseries were detrended. The precipitation index in **a** has been summed over the entire year from July to the following June; in **b**, summed over the high-rainfall months of December–February; and in **c**, measured for October, the onset of the rainy season. The solid lines, at the value of the median for each axis, divide the data into high and low categories. All cases suggest that lower rainfall is associated with El Niño events. Station precipitation data are from the NOAA Baseline Climatological Dataset, compiled by the National Climatic Data Center (Asheville, North Carolina, USA) and the Environmental Research Laboratory (Boulder, Colorado, USA).

climatological monthly mean at each station, we normalized each time series by its standard deviation. We then averaged the four normalized anomaly time series to create an index of central Andes precipitation anomalies. When data were missing (about 9% of the time), we estimated the index as the average of the remaining available station values. Other indices (total rainfall; the first principal component) gave quite similar results.

Figure 3a shows that precipitation anomalies summed over the entire year (July to the following June) are likely to be lower during warmer El Niño years. Precipitation anomalies summed over the highest-precipitation months of December to February show an even tighter relationship (Fig. 3b). October precipitation is also suppressed during El Niño years, suggesting a later onset of the rainy season (Fig. 3c). In keeping with the type of forecasts made by the Andeans, we also divide ('bin') the data into high/low precipitation, and high/low NINO3, and apply χ^2 tests to measure the significance of the relations. (NINO3 is the sea surface temperature anomaly, averaged over 5° S–5° N and 150° W–90° W.) Our ability to test the significance of these relationships is limited by the relatively small amount of data; nevertheless, the test shows high significance (exceeding the 95% level) for the December–February precipitation relationship. The other two relationships shown in Fig. 3 are less significant (around the 75% level).

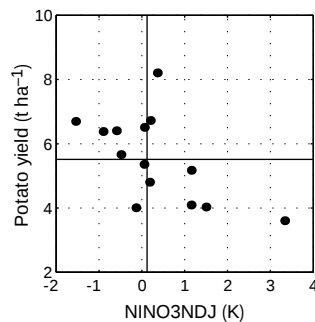


Figure 4 Central Andean harvests and El Niño. Shown are potato yields (in tonnes per hectare) for Puno department plotted against NINO3NDJ, 1980–81 to 1993–94. The solid lines represent median values for each axis.

Given the sensitivity of potatoes to drought, this reduction in precipitation may account for part of the relationship between June high-cloud amount and potato harvests (Fig. 1a). A more complete analysis of the effect of climate on crops must include other climatic variables, most notably temperature. In fact, we find a strong monotonic relationship between ENSO and mean monthly temperatures in the central Andes region, with warmer mean temperatures in the central Andes during El Niño years. Though the effect of temperature on crop growth is not obvious, the consequence of ENSO for potato yields is convincingly demonstrated from preliminary analysis of potato-yield data from Puno department near Lake Titicaca (Fig. 4). Significant reductions in potato yield are observed for warm ENSO years. A 2-bin χ^2 test shows significance at the 90% level. The linear correlation value is -0.6 . □

Received 22 March; accepted 11 November 1999.

- Zimmerer, K. S. *Changing Fortunes: Biodiversity and Peasant Livelihood in the Peruvian Andes* (Univ. California Press, Berkeley, 1996).
- Salomon, F. & Urioste, G. L. *The Huarochiri Manuscript: A Testament of Ancient and Colonial Andean Religion* (Univ. Texas Press, Austin, 1991 [c. 1608]).
- Ballón Aguirre, E., Cerrón Palomino, R. O. & Chambi Apaza, E. *Vocabulario Razonado de la Actividad Agraria Andina. Terminología Agraria Quechua*. (Centro de Estudios Regionales Andinos "Bartolomé de Las Casas", Cusco, 1992).
- Urton, G. *At the Crossroads of the Earth and The Sky: An Andean Cosmology* (Univ. Texas Press, Austin, 1981).
- Grillo Fernández, E., Quiso Choque, V., Rengifo Vásquez, G. & Valladolid Rivera, J. *Crianza Andina de la Chacra* (Proyecto Andino de Tecnología Campesina, Lima, 1994).
- Mishkin, B. Cosmological ideas among the Indians of the southern Andes. *J. Am. Folklore* 53, 225–241 (1940).
- Morote Best, E. La fiesta de San Juan, el Bautista. *Arch. Peruanos Folklore* 1, 160–200 (1953).
- Orlove, B. S. Two rituals and three hypotheses: an examination of solstice divination in southern highland Peru. *Anthropol. Quart.* 52, 86–98 (1979).
- Lira, J. A. *Farmacopea Tradicional Indígena y Prácticas Rituales* (El Condor, Lima, 1946).
- Camino, A., Recharte, J. & Bidegaray, P. *La Tecnología en el Mundo Andino. Runakunap Kawsayninkupaq Rurasankunap* Vol. 1 (eds Lechtman, H. & Soldi, A. M.) 261–281 (Universidad Nacional Autónoma de México, Mexico City, 1981).
- Goland, C. *Cultivating Diversity: Field Scattering as Agricultural Risk Management in Cuyo Cuyo, Dept. of Puno, Peru* (Working Paper No. 4, Research Project on Production, Storage and Exchange in a Terraced Environment on the Eastern Andean Escarpment, Department of Anthropology, Univ. North Carolina, Chapel Hill, 1992).
- Kolata, A. *Valley of the Spirits: A Journey into the Lost Realm of the Aymara* (Wiley, New York, 1996).
- Carter, W. E. & Mamani, M. *Irpa Chico: Individuo y Comunidad en la Cultura Aymara* (Juventud, La Paz, 1982).
- Mamani, M. in *Raíces de América: el Mundo Aymara* (ed. Albó, X.) 75–131 (Alianza, Madrid, 1988).
- Yampara Huarachi, S. in *La Cosmovisión Aymara* (eds van den Berg, H. & Schiffrers, N.) 143–186 (UCB/HISBOL, La Paz, 1992).
- Kraft, K. E. *Andean Fields and Fallow Pastures: Communal Land Use Management under Pressures for Intensification*. Thesis, Univ. Florida (1995).
- Beukema, H. P. & van de Zaag, D. E. *Introduction to Potato Production* (Pudoc, Wageningen, 1990).
- MacKerron, D. K. L. & Jeffries, R. A. The distribution of tuber sizes in droughted and irrigated crops of potato. I. Observations on the effect of water stress on graded yields from differing cultivars. *Potato Res.* 31, 269–278 (1988).
- Bauer, B. S. & Dearborn, D. S. P. *Astronomy and Empire in the Ancient Andes: The Cultural Origins of Inca Sky Watching*. (Univ. Texas Press, Austin, 1995).
- de Arriaga, P. J. *The Extirpation of Idolatry in Peru* (Univ. Kentucky Press, Lexington, 1968 [1621]).
- Anonymous Jesuit priest Misión de las provincias de los Huachos y Yauyos. *Rev. Histórica* (Lima) 6, 180–197 (1919).
- de Santillán, F., Valera, B. & Pachacuti, J. de S. C. *Tres Relaciones de Antigüedades Peruanas* (Guaranía, Asunción, 1950).

- Rossow, W. B. & Schiffer, R. A. ISCCP cloud data products. *Bull. Am. Meteorol. Soc.* 72, 2–20 (1991).
- Kent, G. S. *et al.* Surface temperature related variations in tropical cirrus cloud as measured by SAGE II. *J. Clim.* 8, 2577–2594 (1995).
- Liao, X., Rossow, W. B. & Rind, D. Comparison between SAGE II and ISCCP high-level clouds 1. Global and zonal mean cloud amounts. *J. Geophys. Res.* 100, 1121–1135 (1995).
- Kalnay, E. *et al.* The NCEP/NCAR 40 year reanalysis project. *Bull. Am. Meteorol. Soc.* 77, 437–471 (1996).
- Schaefer, B. E. Astronomy and the limits of vision. *Vistas Astron.* 36, 311–361 (1993).
- Thompson, L. G., Mosley-Thompson, E. & Arnao, B. J. El Niño–Southern Oscillation events recorded in the stratigraphy of the tropical Quelccaya ice cap, Peru. *Science* 226, 50–53 (1984).
- Aceituno, P. On the functioning of the Southern Oscillation in the South American sector. Part I: Surface climate. *Mon. Weath. Rev.* 116, 505–524 (1988).
- Frisancho, A. R. *et al.* Developmental, genetic and environmental components of aerobic capacity at high altitude. *Am. J. Phys. Anthropol.* 96, 431–442 (1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank G. Urton for insights into Andean ethnoastronomy, G. Rasmussen for discussions on visibility and cirrus cloud, G. Scott for providing potato yield data for Puno department, Peru, and R. Bishop, K. Cook, D. Dearborn, D. Helfand, A. Kaplan, G. Kiladis, J. Lenters, M. Sarazin and B. Schafer for comments. M.A.C. thanks B. D'Achille for first making him aware of the Andean forecasting scheme.

Correspondence and requests for materials should be addressed to B.S.O. (e-mail: bsorlove@ucdavis.edu).

Channelized fluid flow in oceanic crust reconciles heat-flow and permeability data

A. T. Fisher* & K. Becker†

* Earth Sciences Department, University of California, Santa Cruz, California 95064, USA

† Division of Marine Geology and Geophysics, Rosenstiel School of Marine and Atmospheric Sciences, University of Miami, Miami, Florida 33145, USA

Hydrothermal fluid circulation within the sea floor profoundly influences the physical, chemical and biological state of the crust and the oceans. Circulation within ridge flanks (in crust more than 1 Myr old) results in greater heat loss^{1–3} and fluid flux⁴ than that at ridge crests and persists for millions of years, thereby altering the composition of the crust and overlying ocean^{5,6}. Fluid flow in oceanic crust is, however, limited by the extent and nature of the rock's permeability⁷. Here we demonstrate that the global data set of borehole permeability measurements in uppermost oceanic crust^{7–9} defines a trend with age that is consistent with changes in seismic velocity^{10,11}. This trend—which indicates that fluid flow should be greatly reduced in crust older than a few million years—would appear to be inconsistent with heat-flow observations, which on average indicate significant advective heat loss in crust up to 65 Myr old³. But our calculations, based on a lateral flow model, suggest that regional-scale permeabilities are much higher than have been measured in boreholes. These results can be reconciled if most of the fluid flow in the upper crust is channelized through a small volume of rock, influencing the geometry of convection and the nature of fluid–rock interaction.

Permeabilities have been measured in oceanic crustal boreholes using a drill-string packer, a device that hydraulically seals a borehole so that fluid can be pumped into the formation and transient changes in pressure can be monitored⁷. These measurements yield “bulk” permeabilities close to the borehole, assumed to be distributed evenly within the isolated interval. Other methods, including interpretation of temperature logs within upper basement^{12,13}, and modelling of subsurface tidal response recorded by pressures in

long-term sealed observatories¹⁴, yield permeability estimates at larger scales. Until recently, borehole measurements were restricted to crustal ages ≥ 6 Myr and did not indicate any permeability evolution trend; instead, these data suggested a narrow range for uppermost oceanic basement, 10^{-13} to 10^{-14} m² (ref. 7).

New borehole measurements in 0.9–3.6 Myr-old crust on the eastern flank of Juan de Fuca ridge⁸, in combination with measurements made elsewhere^{7,9}, reveal a remarkably consistent trend in permeability evolution within uppermost basaltic basement (Fig. 1). Bulk permeability decreases rapidly as young crust ages, from about 10^{-10} m² to 10^{-13} m² over the first 3–4 Myr (Fig. 1a and b).

Recent compilations and surveys of seismic velocities in uppermost oceanic crust^{10,11} show a similar trend: velocity increases rapidly during the first few million years, then more slowly during later times (Fig. 1c and d). Density calculations from near-bottom gravity measurements¹⁵ also show a rapid increase in the first 1 Myr, but wireline logs in older boreholes suggest only small additional increases with age. Thus physical properties measurements are consistent with initially rapid pore infilling by mechanical and magmatic processes during the earliest stages of crustal evolution, and by slower diagenetic processes as ageing continues^{16,17}.

Global heat flow compilations³ appear to conflict with this rapid-ageing model because they suggest that advective heat loss continues on average out to 65 Myr (Fig. 1e and f), despite the early permeability reduction. Regional heat-flow surveys indicate continued circulation in basement to even greater ages^{18,19}. The global heat-

flow data also show little dependence on sediment thickness, suggesting that a process other than sediment thickening may limit advective heat loss from the crust³. Here we suggest a model by which the physical properties and heat-flow data sets can be reconciled, based on lateral fluid fluxes required by the heat-flow data, calculations of fluid driving forces, and inferences regarding crustal heterogeneity.

We assume that the global heat-flow data set reliably indicates advective heat loss from the crust (Fig. 1e, inset), and that hydrothermal circulation locally homogenizes temperatures in uppermost basement. Cold sea water enters this “well-mixed aquifer” and warms while flowing laterally until discharged at the sea floor²⁰ (Fig. 2a). Heat flow through sediments is mainly conductive, as vertical fluid flow is greatly restricted once the sediments are a few tens of metres thick. Net lateral fluid flow is the critical geometry for widespread heat removal from basement, because local convection will only redistribute heat, and vertical flow is inefficient at extracting heat over long horizontal distances. Vigorous flow in the uppermost crust is consistent with the global permeability data set^{7,8} and with evidence for high water/rock ratios during alteration^{5,21}.

Heat is ‘mined’ by laterally flowing fluid, with heat flow suppression (q_m/q_L , where q_m is measured heat flow and q_L is lithospheric heat flow) dependent on sediment thickness and the basement layer thickness in which lateral flow occurs (h_s and h_b , respectively), sediment thermal conductivity (λ_s), the density and specific heat of water (ρ and c , respectively), distance from the fluid entry point (Δx), and the specific discharge of fluid in basement (Q_0 , the volume flux per cross-sectional area)²⁰:

$$\frac{q_m}{q_L} = \left[1 - \exp\left(\frac{-\lambda_s \Delta x}{Q_0 h_s h_b \rho c}\right) \right] \quad (1)$$

Sea-floor heat flow suppression as a function of age is known (Fig. 1e, inset), and we estimate sediment thickness based on crustal age and sedimentation rate, let Δx vary with crustal age, and calculate Q_0 .

The driving forces moving fluid laterally through the crust can be linked to equation (1) using a form of Darcy’s law:

$$Q_0 = -\frac{k \, dP}{\mu \, dx} \quad (2)$$

Here μ is fluid viscosity, P is pressure, and k is formation permeability. In contrast to the situation at the ridge crests, there is only one significant source of lateral pressure gradients within ridge flanks: the difference between cooler and warmer columns of fluid entering and leaving the crust. The greatest difference in fluid pressure would arise from a connection between a cool recharge zone, in which bottom sea water moves rapidly into the crust, and a warm discharge zone, in which heated hydrothermal water rises adiabatically into the overlying ocean.

Combining equations (1) and (2) allows calculation of the formation permeability of upper oceanic crust required to allow sufficient lateral flow so as to explain the observed heat-flow suppression. We include a temperature dependence in fluid viscosity and density, and assume that the mean distance between fluid entry and exit points increases with time, because faults and basement outcrops that extend to the sea floor are likely to become more widely spaced with age as sediment thickens. dP/dx is calculated from the difference in fluid density between downflow and upflow zones, and by letting $dx = 5.0 + 0.5 \times \text{age}$ (in km and Myr, respectively), which yields distances between upflow and downflow zones of 5.0–37.5 km over ages of 1–65 Myr. We let Δx , the characteristic lateral flow length, be one half the downflow–upflow spacing, consistent with apparent scales of flow in several settings^{20,22–24}. The temperature assumed for the fluid within upflow zones is that at the sediment–basement interface, with lithospheric heat flow based on crustal age, and the pressure difference between downflow and upflow zones is calculated as

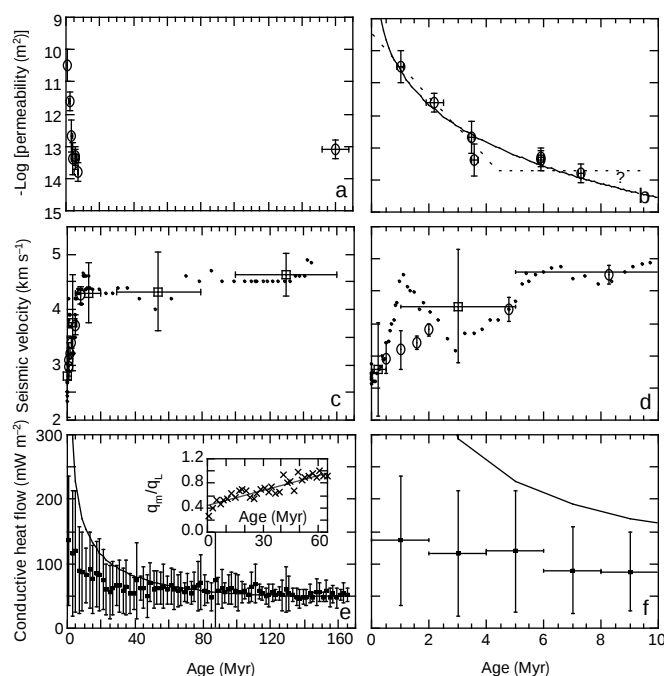


Figure 1 Bulk properties within uppermost oceanic crust, and heat loss versus crustal age. **a**, Permeability for 0–165 Myr (refs 7, 8). Horizontal and vertical bars indicate estimated uncertainties in crustal age and bulk permeability. **b**, Permeability for 0–10 Myr. Curves are least-squares best-fit exponential (solid) and straight-line segments fitted by eye (dashed). The ? indicates that the trend beyond 6–7 Myr is not constrained by observations. **c**, Seismic velocity for 0–165 Myr. Dots are medians of a moving 9-point window of a global data set¹⁰; squares are means, and 1 s.d. (vertical bars) and range of interval (horizontal bars) are also shown for same data; open circles are results from the flank of the East Pacific Rise¹¹, with 1 s.d. (vertical bars). **d**, Seismic velocity for 0–10 Myr. **e**, Oceanic heat flow and predicted values (GDH³) based on crustal age, using 2-Myr bins (horizontal bars) and showing 1 s.d. within each bin (vertical bars). Inset, heat-flow suppression as a function of crustal age, with a least-squares best-fit line of $q_m/q_L = 0.43 + 0.0085 \times \text{age}$ (age in Myr). **f**, Sea-floor heat flow versus age for 0–10 Myr.

$dP \approx \Delta \rho g(h_s + h_b)$. Because there are practical limits on the available driving forces (described below), and because crustal permeability varies by many orders of magnitude, these calculations are robust to reasonable modifications to the stated assumptions.

The maximum driving forces available for lateral fluid flow within young ridge flanks (1–5 Myr) are a few kilopascals within

crust having a thin sediment layer, to several hundred kilopascals if the sediment layer is thicker (Fig. 2b), because permeable basement is deeper and hotter. Pressure differences increase more rapidly with age within crust experiencing higher sedimentation rates. We calculate higher pressure differences for young crust if we use the temperature at the base of the permeable layer to calculate the upflow temperature, but these pressure differences decrease with time because the effect of a rapid decline in lithospheric heat input initially overwhelms the effect of increasing sediment thickness. The lateral specific discharge required to suppress sea-floor heat flow decreases as the crust ages (Fig. 2c), because a deeper hydrothermal layer is hotter and can mine heat more efficiently, and because the thermal anomaly decreases with age (Fig. 1e and f). For young ridge flanks, these calculations indicate specific discharge of the order of 10^{-8} to 10^{-6} m s^{-1} ($0.3\text{--}30 \text{ m yr}^{-1}$), consistent with other models^{20,22,24} and with ^{14}C age estimates for crustal fluids²⁵.

The calculated formation-scale permeabilities vary with sediment thickness and flow depth, and the lowest permeabilities correspond to the greatest available pressure differences (Fig. 2d). Calculated permeabilities decrease with crustal age, but overlap borehole observations only for very young crust ($\leq 2 \text{ Myr}$) having a thick permeable layer (500 m) and a high sedimentation rate (50 m Myr^{-1}) (Fig. 2d). None of the calculated permeability evolution curves include the abrupt decrease with age apparent in the borehole data for young crust, and calculations based on more typical abyssal sedimentation rates (5 m Myr^{-1}) are offset from observations by several orders of magnitude (Fig. 2d). The lowest calculated permeabilities (10^{-14} to 10^{-13} m^2 at 50–65 Myr) correspond to the largest available driving forces (1.0–3.0 MPa), which would be greater than lithostatic overburden unless sediment is very thick. Recent measurements from ridge-flank borehole observatories²⁶ indicate more modest differences in crustal fluid pressures, a few to several tens of kilopascals, as do estimates of flow rates from open hole temperature logs¹³.

Matching the observed borehole permeability evolution trend would require a combination of deepening or thickening of the hydrothermal layer with age, or upflow and downflow zones that move closer together, which seems unlikely. We offer an alternative explanation. The bulk properties of upper oceanic crust (seismic, density and permeability) are correct as measured. Very young crust is fractured and open, seismic velocities are low, and bulk permeability is high. Within a few million years after leaving the ridge axis, much of the crustal porosity is lost, density and seismic velocity increase, and bulk permeability decreases. Large-scale hydrothermal circulation responsible for the documented thermal anomalies continues within discrete channels, spatially restricted regions that occupy a small volume of the crust. This concept is analogous to flow channelling within fractures and fault planes, a process known to control solute transport and water–rock interaction within fractured aquifers at a smaller scale²⁷. In the upper oceanic crust, the extrusive section would contain the primary ‘flow plane’, with flow occurring preferentially through a network of connected channels such as breccia zones and pillow and flow boundaries. These channels would take the longest to infill, but they would have little effect on seismic velocities or density if they are metres or tens of metres apart and occupy a small crustal volume. Because thermal conduction is efficient, sea-floor heat-flow suppression would be indistinguishable from that resulting from ubiquitous flow within a more homogeneous upper basement.

Outcrops and high-angle normal faults would be critical to this channelized flow system, allowing fluids to penetrate through lower-permeability sediments and basement. The driving forces for large-scale flow could remain modest (Fig. 2b), as large-scale permeabilities would be high, and the flow channels would remain open until the crust is old because of their chemical isolation. At an average age of 65 Myr, continuing diagenesis, lower lithospheric heat flow, and increased spacing between outcrops and sediment-

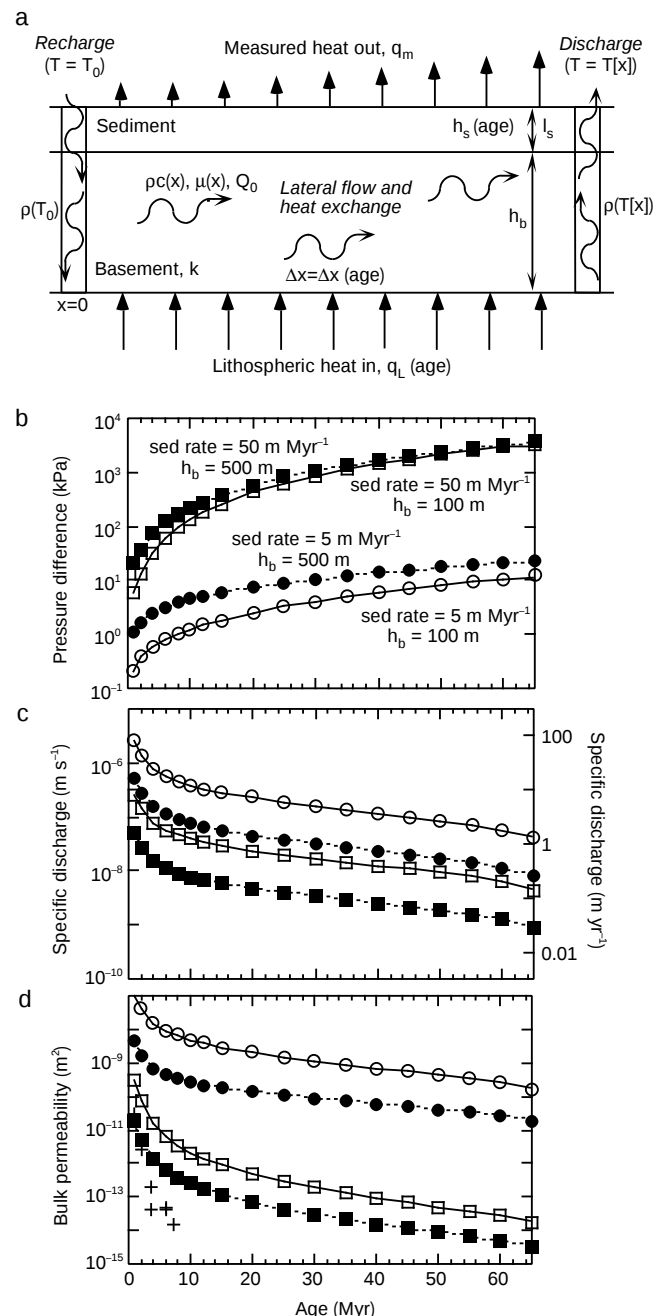


Figure 2 Conceptual model and calculations for lateral fluid flow through upper oceanic crust. **a**, Diagram of the ‘well-mixed aquifer’ model (modified from ref. 20). Symbols defined in text. Some parameters are labelled on plots; other values are: $\lambda_s = 1.2 \text{ W m}^{-1} \text{ K}^{-1}$, $T_0 = 0^\circ \text{C}$, $c = 4,200 \text{ J kg}^{-1} \text{ K}^{-1}$, ρ and μ are calculated using the mean temperature along the flow path, h_s and Δx vary with crustal age (as described in text). **b**, Calculated pressure difference between recharge and discharge locations, assuming adiabatic vertical flow. Crustal parameters as labelled. **c**, Specific discharge (volume flux per area) for crustal parameters as in Fig. 1b, based on the linear fit to global heat-flow data (Fig. 1e, inset). **d**, Apparent bulk permeability for crustal parameters as in Fig. 1b, required by the observed heat-flow suppression (Fig. 1e), and necessary to allow the calculated specific discharge, given the largest available pressure gradient. Borehole permeability values plotted for reference (crosses).

penetrating faults would make lateral fluid flow within upper basement inefficient at removing heat. Circulation would continue within basement locally, redistributing heat or being thermally masked below thick sediments, but heat loss from the crust would be entirely conductive.

This model explains how free convection could be prevented despite very high effective permeabilities, enhancing the efficiency of lateral flow for advecting heat. It does not preclude local thermal homogenization of upper basement²⁸, either by organized or chaotic flow²⁹, but local convection could also be channelized. Unlike a homogeneous permeability model for the upper crust, our model is consistent with observed lithostratigraphic, hydrologic, and alteration heterogeneity^{5,7,17,30}. There would still be chemically and biologically significant fluid flow within much of the upper crust, but fluid fluxes away from the primary channels would be small relative to those in the channels.

Some important implications for hydrogeology and fluid–rock interaction within oceanic ridge flanks follow from our model. First, most of the fluid flow within the uppermost oceanic crust of ridge flanks would be very strongly guided (perhaps fully constrained) by the permeability distribution. There would be no tendency for hydrothermal convection cells to form with any particular geometry, and flow channelling would reduce the tendency for small-scale convection cells to form, since crustal permeability would be both heterogeneous and highly anisotropic. Second, flow would be much less chaotic than in a homogeneous, isotropic porous system having the same formation-scale permeability. Local flow paths would be strongly influenced by crustal constructional, tectonic and alteration patterns. Third, water–rock ratios would be highly variable over small spatial distances, depending on proximity to primary flow channels, and reaction in surrounding rock would be limited by slower advection and diffusion away from the channels. Borehole permeability estimates from packer testing would indicate the bulk hydrologic properties of most of the upper crust—with decreasing permeability associated with infilling of pores and increasing seismic velocities in young crust—but the flow channels responsible for the ridge–flank heat-flow deficit (Fig. 1e), being spatially rare, would generally not be penetrated by vertical boreholes⁸.

Testing of the model we report here can be performed by continuing to map out bulk permeability, apparent water age, and the distribution of lateral and vertical pressure gradients within oceanic crust, over a range of crustal ages. Understanding will also come from measuring permeability at a variety of lateral scales using the same boreholes and the same measurement methods, and by cross-hole testing, in order to establish the spatial dependence of properties. Coupled heat and fluid flow models will continue to provide insight, but should include heterogeneous permeability distributions and anisotropy. □

Received 18 May; accepted 11 November 1999.

1. Sclater, J. G., Jaupart, C. & Galson, D. The heat flow through oceanic and continental crust and the heat loss of the earth. *Rev. Geophys. Space Phys.* **18**, 269–311 (1980).
2. Stein, C. & Stein, S. A model for the global variation in oceanic depth and heat flow with lithospheric age. *Nature* **359**, 123–129 (1992).
3. Stein, C. & Stein, S. Constraints on hydrothermal heat flux through the oceanic lithosphere from global heat flow. *J. Geophys. Res.* **99**, 3081–3095 (1994).
4. Mottl, M. J. & Wheat, C. G. Hydrothermal circulation through mid-ocean ridge flanks: fluxes of heat and magnesium. *Geochim. Cosmochim. Acta* **58**, 2225–2237 (1994).
5. Alt, J. C. in *Seafloor Hydrothermal Systems: Physical, Chemical, Biological and Geological Interactions* (eds Humphris, S. E., Zierenberg, R. A., Mullineaux, L. S. & Thompson, R. E.) 85–114 (American Geophysical Union, Washington DC, 1995).
6. Elderfield, H. & Schultz, A. Mid-ocean ridge hydrothermal fluxes and the chemical composition of the ocean. *Annu. Rev. Earth Planet. Sci.* **24**, 191–224 (1996).
7. Fisher, A. T. Permeability within basaltic oceanic crust. *Rev. Geophys.* **36**, 143–182 (1998).
8. Becker, K. & Fisher, A. T. Permeability of upper oceanic basement on the eastern flank of the Endeavor Ridge determined with drill-string packer measurements. *J. Geophys. Res.* (in the press).
9. Bruns, T. & Lavoie, D. Bulk permeability of young backarc basalt in the Lau Basin from a downhole packer experiment (Hole 839B). *Proc. ODP Sci. Res.* **135**, 805–816 (1994).
10. Carlson, R. L. Seismic velocities in the uppermost oceanic crust: age dependence and the fate of layer 2A. *J. Geophys. Res.* **103**, 7069–7077 (1998).

11. Grevenmeyer, I., Norbert, K., Villinger, H. & Weigel, W. Hydrothermal activity and the evolution of the seismic properties of upper oceanic crust. *J. Geophys. Res.* **104**, 5069–5079 (1999).
12. Becker, K., Langseth, M., Von Herzen, R. P. & Anderson, R. Deep crustal geothermal measurements, Hole 504B, Costa Rica Rift. *J. Geophys. Res.* **88**, 3447–3457 (1983).
13. Fisher, A. T., Becker, K. & Davis, E. E. The permeability of young oceanic crust east of Juan de Fuca Ridge determined using borehole thermal measurements. *Geophys. Res. Lett.* **24**, 1311–1314 (1997).
14. Wang, K. & Davis, E. E. Theory for the propagation of tidally induced pore pressure variations in layered subseafloor formations. *J. Geophys. Res.* **101**, 11483–11495 (1996).
15. Holmes, M. L. & Johnson, H. P. Upper crustal densities derived from seafloor gravity measurements: northern Juan de Fuca Ridge. *Geophys. Res. Lett.* **17**, 1871–1874 (1993).
16. Jacobson, R. S. Impact of crustal evolution on changes of the seismic properties of the uppermost oceanic crust. *Rev. Geophys.* **30**, 23–42 (1992).
17. Gillis, K. M. & Sapp, K. Distribution of porosity in a section of upper oceanic crust exposed in the troodos ophiolite. *J. Geophys. Res.* **102**, 10133–10149 (1997).
18. Embley, R., Hobart, M., Anderson, R. & Abbott, D. Anomalous heat flow in the northwest Atlantic, a case for continued hydrothermal circulation in 80 MY crust. *J. Geophys. Res.* **88**, 1067–1074 (1983).
19. Noel, M. & Hounslow, M. W. Heat flow evidence for hydrothermal convection in Cretaceous crust of the Madeira Abyssal Plain. *Earth. Planet. Sci. Lett.* **90**, 77–86 (1988).
20. Langseth, M. G. & Herman, B. Heat transfer in the oceanic crust of the Brazil Basin. *J. Geophys. Res.* **86**, 10805–10819 (1981).
21. Gillis, K. & Robinson, P. T. Distribution of alteration zones in the upper oceanic crust. *Geology* **16**, 262–266 (1988).
22. Baker, P., Stout, P., Kastner, M. & Elderfield, H. Large-scale lateral advection of seawater through oceanic crust in the central equatorial Pacific. *Earth. Planet. Sci. Lett.* **105**, 522–533 (1991).
23. Langseth, M. G., Becker, K., Von Herzen, R. P. & Schultheiss, P. Heat and fluid flux through sediment on the western flank of the Mid-Atlantic Ridge: a hydrogeological study of North Pond. *Geophys. Res. Lett.* **19**, 517–520 (1992).
24. Davis, E. E. *et al.* Regional heat-flow variations across the sedimented Juan de Fuca Ridge eastern flank: constraints on lithospheric cooling and lateral hydrothermal heat transport. *J. Geophys. Res.* **104**, 17675–17688 (1999).
25. Elderfield, H., Wheat, C. G., Mottl, M. J., Monnin, C. & Spiro, B. Fluid and geochemical transport through oceanic crust: a transect across the eastern flank of the Juan de Fuca Ridge. *Earth. Planet. Sci. Lett.* **172**, 151–165 (1999).
26. Davis, E. & Becker, K. Borehole observations record driving forces for hydrothermal circulation in young oceanic crust. *Eos* **79**, 369F377–378 (1998).
27. Tsang, C. F. & Neretnieks, I. Flow channeling in heterogeneous fractured rocks. *Rev. Geophys.* **36**, 275–298 (1998).
28. Davis, E. E., Chapman, D. S., Forster, C. & Villinger, H. Heat-flow variations correlated with buried basement topography on the Juan de Fuca Ridge flank. *Nature* **342**, 533–537 (1989).
29. Davis, E. E. *et al.* An unequivocal case for high Nusselt-number hydrothermal convection in sediment-buried igneous oceanic crust. *Earth. Planet. Sci. Lett.* **146**, 137–150 (1997).
30. Matthews, M., Salisbury, M. & Hyndman, R. Basement logging on the Mid-Atlantic Ridge, Deep Sea Drilling Project Hole 3958. *Init. Rep. DSDP 78B*, 717–730 (1984).

Acknowledgements

This work was supported by grants from the National Science Foundation and the United States Science Support Program to the Ocean Drilling Program. We thank C. Stein for suggestions, and P. Stauffer, J. Stein and E. Giambalvo for discussions that improved the manuscript.

Correspondence and requests for materials should be addressed to A.T.F. (e-mail: afisher@es.ucsc.edu).

A highly unsaturated fatty acid predicts carbon transfer between primary producers and consumers

Dörthe C. Müller-Navarra*, Michael T. Brett†, Anne M. Liston* & Charles R. Goldman*

* Department of Environmental Science and Policy, University of California, Davis, California 95616, USA

† Department of Civil & Environmental Engineering, Box 352700, University of Washington, Seattle, Washington 98195, USA

The factors that regulate energy transfer between primary producers and consumers in aquatic ecosystems have been investigated for more than 50 years (refs 1–3). Among all levels of the food web (plants, herbivores, carnivores), the plant–animal interface is the most variable and least predictable link^{4–6}. In hyper-eutrophic lakes, for example, biomass and energy transfer is often inhibited at the phytoplankton–zooplankton link⁴, resulting in an accumulation of phytoplankton biomass instead of sustaining

production at higher trophic levels, such as fish. Accumulation of phytoplankton (especially cyanobacteria) results in severe deterioration of water quality, with detrimental effects on the health of humans and domestic animals, and diminished recreational value of water bodies^{7,8}. We show here that low transfer efficiencies between primary producers and consumers during cyanobacteria bloom conditions are related to low relative eicosapentaenoic acid (20:5 ω 3) content of the primary producer community. Zooplankton growth and egg production were strongly related to the primary producer 20:5 ω 3 to carbon ratio. This indicates that limitation of zooplankton production by this essential fatty acid is of central importance at the pelagic producer–consumer interface.

During the summer and winter 1997 and the spring of 1998, we performed a series of laboratory experiments in which seston (fine particles in water) collected from hypereutrophic Stonegate pond was fed to the herbivorous zooplankter *Daphnia magna* as a consumer. Trophic transfer was measured as *Daphnia* biomass accrual. The experiments were performed at a constant temperature using standardized animals in a flow-through experimental system. This enabled us to determine directly the effect of phytoplankton elemental and biochemical composition on energy transfer to herbivore biomass, independent of factors such as temperature and predation which can also influence this process⁹.

Stonegate pond is a hypereutrophic (mean total phosphorus, $125 \pm 34 \mu\text{g l}^{-1}$) shallow pond situated in Davis, California, USA. During the summer, cyanobacteria (mainly *Oscillatoria* c.f. *tenuis*, and *Anabaena* c.f. *aphanizomenoides*) strongly dominated the phytoplankton community. Although seston carbon was quite high (3.9–9.4 mg C l⁻¹), experimentally determined growth rates for *Daphnia* feeding on the summer phytoplankton assemblage were very low: around the threshold for egg production (0–0.5 eggs per clutch). The food quality of seston was poor and the transfer efficiency between primary producers and consumers was low, between 5% and 26% of consumer ingested carbon. During the winter and spring, the phytoplankton community was dominated by diatoms. Although the seston carbon concentration was lower (1.7–3.8 mg C l⁻¹) than in the summer, daphnids grew at about their maximal rate at the temperature of our experiments and produced large first clutches (9.1–17.0 eggs per clutch). Conversion of consumed phytoplankton carbon into herbivore biomass was 50–65%. *Daphnia* growth rates and carbon concentrations were in fact inversely related (Table 1), showing that food quality, rather than quantity, regulated carbon transfer between primary producers and consumers in this system.

We measured a variety of chemical characteristics (chlorophyll *a* (chl *a*) content, elemental and fatty acid composition) of the

Table 1 Regressions with *Daphnia* growth rates

Independent variable	Slope	Y-intercept	<i>P</i>	<i>r</i> ²
C (mg l ⁻¹)	-0.05	0.52	0.002	0.55
P/C (mol mol ⁻¹)	27.11	-0.005	0.0087	0.45
N/P (mol mol ⁻¹)	-0.013	0.48	0.0031	0.53
N/C (mol mol ⁻¹)	-4.39	1.04	<0.0001	0.85
Chl- <i>a</i> /C ⁻¹ (μg mg ⁻¹)	-0.01	0.37	0.445	0.05
18:3 ω 3/C (μg mg ⁻¹)	-0.02	0.42	0.063	0.26
18:4 ω 3/C (μg mg ⁻¹)	0.14	0.04	<0.0001	0.80
20:5 ω 3/C (μg mg ⁻¹)	0.11	0.04	<0.0001	0.95
22:6 ω 3/C (μg mg ⁻¹)	0.31	0.13	0.008	0.50
$\Sigma\omega$ 3-PUFA/C (μg mg ⁻¹)	3.83	0.11	0.3761	0.07

Chl-*a*, chlorophyll *a*; 18:3 ω 3, linolenic acid; 18:4 ω 3, octadecatetraenoic acid; 20:5 ω 3, eicosapentaenoic acid; 22:6 ω 3, docosahexaenoic acid; PUFA, polyunsaturated fatty acids.

phytoplankton community, and related these to daphnid growth rates. Regression analyses were performed to determine which of these characteristics best predicted biomass production by these herbivorous zooplankters. The chl-*a*/C ratio was not significantly related to *Daphnia* growth (Fig. 1, Table 1). Chlorophyll *a* is often used as a food parameter because it allows one to infer the relative contributions of phytoplankton and presumably low-quality detritus to the seston C pool. However, our results suggest that the chl-*a*/C ratio may not be a particularly useful food-quality index. The molar elemental P/C ratio was moderately positively (that is, molar C/P was inversely) correlated to daphnid growth in these experiments (Fig. 2, Table 1). However, minimal P/C (maximal C/P) molar ratios observed were higher than 4.7×10^{-3} (less than 210), and most researchers believe phosphorus limitation of *Daphnia* growth will only occur at P/C ratios below 3.3×10^{-3} (C/P ratios above 300)^{10,11}. The seston N/P ratio was also correlated to the daphnid growth rates (Table 1), which might well be an indirect effect due to correlation between seston N/P and its 20:5 ω 3 content ($P = 0.0005$, $r^2 = 65$). Counter-intuitively, the molar N/C ratio was negatively related to daphnid growth rates (Table 1), indicating that growth was negatively related to the nitrogen moiety.

In contrast to these parameters, we obtained a very strong correlation between the seston 20:5 ω 3 to carbon (20:5 ω 3/C) ratio and *Daphnia* growth rates (Fig. 3, Table 1) and egg production (Fig. 4). The 20:5 ω 3/C ratio was clearly the best predictor of *Daphnia* growth of fatty acids measured (Table 1). The fit with other fatty acids was substantially weaker, but still strong (Table 1). However, this is probably due to an inter-correlation with 20:5 ω 3 (for example, $P = 0.0002$, $r^2 = 0.70$ for 18:4 ω 3). Variability in phytoplankton 20:5 ω 3 content could explain the poor predictability of energy transfer at the primary producer–consumer interface. The

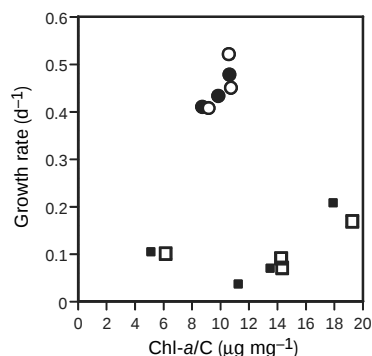


Figure 1 The relationship between the seston chlorophyll *a* to carbon ratio and daphnid growth rates. Squares denote the cyanobacteria-dominated summer assemblages, and circles denote the diatom- and cryptophyte-dominated winter/spring assemblages. Open symbols represent the seston fraction less than 30 μm , and closed symbols the total seston.

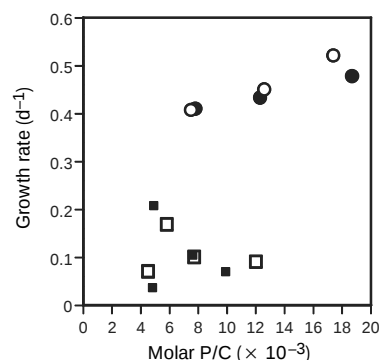


Figure 2 The relationship between the seston molar phosphorus to carbon ratio and daphnid growth rates. Symbols as in Fig. 1.

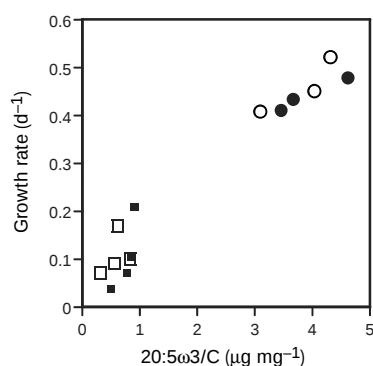


Figure 3 The relationship between the seston 20:5ω3 to carbon ratio and daphnid growth rates. Growth rate = $0.74 [1 - \exp(-0.25x + 0.01)]$, where x is the ratio 20:5ω3/C, $r^2 = 0.96$; for linear regression see Table 1. Symbols as in Fig. 1.

phytoplankton's relative content of this highly unsaturated fatty acid was a strong predictor of zooplankton carbon conversion (that is, secondary production) in this hypereutrophic system. In contrast to the present results, the absolute concentration of 20:5ω3 (that is, 10:5ω3 per litre) was found to be the best predictor of *Daphnia* growth when fed seston from a mesotrophic lake (Schöhsee, Germany)¹². Unlike mesotrophic Schöhsee, where food carbon concentrations were below saturating *Daphnia* growth, seston concentrations in Stonegate pond were well above the growth saturation level. Thus, 20:5ω3 seems to be of general importance for the trophic transfer of energy and elements within the aquatic food web¹³, both in an absolute sense at low food concentrations and in a relative sense at high seston concentrations.

In Stonegate pond, the 20:5ω3 was diluted by the vast amount of carbon present during summer cyanobacteria dominance. Prokaryotic cyanobacteria have a fatty-acid pattern distinct from eukaryotes and generally do not contain highly unsaturated fatty acids like 20:5ω3 (refs 14–16). In contrast, cryptophytes and especially diatoms contain high amounts of 20:5ω3 (refs 14–16). The source of 20:5ω3 under cyanobacteria-dominated conditions was probably cryptophytes, which are often present in low numbers during cyanobacteria blooms. Although relatively rare, 20:5ω3-rich algal taxa might be of prime importance for trophic transfer at the producer–consumer interface during cyanobacteria-dominated conditions.

Cyanobacteria have long been known to be of poor quality for herbivorous zooplankton^{17–19}, but the causes of this low food quality are still poorly understood: mechanical interference, toxicity and nutritional inadequacy have been suggested^{7,8,17–20}. Although experimental evidence is contradictory, mechanical interference of the filtering process is often thought to be the main reason for poor food quality of filamentous cyanobacteria, allowing them to build up nuisance blooms in nutrient-enriched lakes. Despite the fact that filamentous cyanobacteria dominated the phytoplankton assemblage during the summer, daphnid growth rates in the total seston fraction were not significantly depressed relative to the less than 30-μm size seston fraction ($F_{1,9} = 2.37$, $P = 0.1638$, analysis of covariance), although large interfering cyanobacteria filaments were more prevalent in the total seston fraction. In contrast to the prevailing view of mechanical interference, our results show that the lack of 20:5ω3 in cyanobacteria could itself be responsible for the poor quality of seston dominated by cyanobacteria. The lack of 20:5ω3 in cyanobacteria appears to be the main reason for the decoupling between primary and secondary production. This is a well known problem when trying to biomanipulate the food web with the aim of reducing phytoplankton biomass by altering higher trophic levels⁷. Our findings may therefore be relevant to lake restoration measures that utilize biomanipulation approaches. □

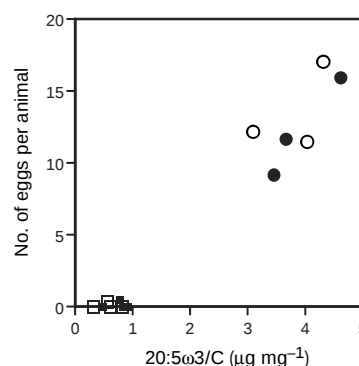


Figure 4 The relationship between the seston 20:5ω3 to carbon ratio and mean number of eggs produced by daphnids (mean clutch size). No. of eggs per animal = $4.0x - 2.5$, where x is the ratio 20:5ω3/C; $P < 0.00001$; $r^2 = 0.96$. Symbols as in Fig. 1.

Methods

Experiments

We performed seven laboratory growth experiments with two treatments (total seston, <30-μm seston) run in duplicate. Growth rates were determined as $g = (\ln W_t - \ln W_0)/t$ where W is the mean body weight of the experimental animals ($n = 8$) at the beginning (W_0) and end (W_t) of the experimental duration of $t = 4$ (± 0.1 d, resulting from 2 h required to process the animals). The carbon transfer efficiency was calculated as the ratio of daphnid carbon accrual to carbon ingested. Daphnid carbon was assumed to be 40% of the dry weight of the animals, as determined from a sub-sample of the experimental animals. Hourly ingestion rates were assumed to be 3.2% of the carbon weight of the daphnids²¹. Daphnids were grown in a flow-through system²² which minimized the influence of grazing and sedimentation on food (seston) concentrations. This ensured that seston was maintained constant at the ambient level. The entire flow-through system was maintained at 20 ± 0.1 °C in a temperature-controlled water bath in a temperature-controlled room. The water was pumped from continuously stirred reservoirs by a multichannel peristaltic pump (Watson-Marlow) to individual 250-ml chambers at the rate of 1.61 per day per chamber. Fresh water from Stonegate pond was collected daily, and ambient large zooplankton were removed before feeding if necessary. *Daphnia magna* used in these experiments originated from a single clone, and were raised under standardized conditions. At the start of each experiment, eight 3-day-old daphnids were pipetted into each chamber. Juveniles were selected randomly from offspring released over 8 h and raised with *Scenedesmus acutus* for 3 d until the start of the experiment. At the end of each experiment, all 7-day-old animals were collected, examined for body length and clutch size under a dissecting microscope, dried and weighed.

Analyses

Particulate matter was collected in duplicate. Total seston was filtered directly onto pre-combusted (24 h at 450 °C) glass-fibre filters (Whatman GF/C). Seston for the <30-μm fraction was screened three times through a 30-μm mesh and then filtered as above. Seston carbon and nitrogen content was determined with a CHN analyser (Perkin-Elmer Model 2400). Particulate phosphorus was calculated by subtracting soluble reactive phosphorus (SRP) from total phosphorus (TP). Water for SRP analyses was filtered through a glass-fibre filter and then measured colorimetrically²³. TP was analysed in unfiltered water by colorimetric means after autoclaving²³. Chl-*a* concentrations were determined using a fluorometric method with acid correction for degradation products²⁴. Fatty acids were analysed, after extraction and methylation²⁵, with a gas chromatograph (HP 6890) equipped with a programmable temperature vaporizer-injector, a fused silica DB-WAX (J&W Scientific) capillary column and a flame ionization detector.

Received 17 May; accepted 31 October 1999.

- Lindemann, R. L. The trophic-dynamic aspect of ecology. *Ecology* **23**, 399–418 (1942).
- Lampert, W. (ed.) In Food limitation and the structure of zooplankton communities. *Arch. Hydrobiol. Beih. Ergebn. Limnol.* **21**, V–VII (1985).
- Gulati, R. D. & Demott, W. (eds) The role of food quality for zooplankton. *Freshwat. Biol.* **38**, (1997).
- McQueen, D. J., Johannes, M. R. S., Post, J. R., Stewart, T. J. & Lean, D. R. S. Bottom-up and top-down impacts on freshwater pelagic community structure. *Ecol. Monogr.* **59**, 289–309 (1989).
- Brett, M. T. & Goldman, C. R. A meta-analysis of the freshwater trophic cascade. *Proc. Nat. Acad. Sci. USA* **93**, 7723–7726 (1996).
- Brett, M. T. & Goldman, C. R. Consumer versus resource control in freshwater pelagic food webs. *Science* **275**, 384–386 (1997).
- Gulati, R. D., Lammens, E. H. R. R., Meijer, M.-L. & Van Donk, E. (eds) *Biomanipulation—Tool for Water Management Hydrobiologia* **200/201**, 29–41 (1990).
- Carmichael, W. W. The toxins of cyanobacteria. *Sci. Am.* **270**, 64–72 (1994).
- Müller-Navarra, D. C. & Lampert, W. Seasonal patterns of food limitation in *Daphnia galeata*: separating food quantity and quality effects. *J. Plankt. Res.* **18**, 1137–1157 (1996).
- Urabe, J. & Watanabe, Y. Possibility of N or P limitation for planktonic cladocerans: an experimental test. *Limnol. Oceanogr.* **37**, 244–251 (1992).

11. Sterner, R. W. *Daphnia* growth on varying quality of *Scenedesmus*: mineral limitation of zooplankton. *Ecology* **74**, 2351–2360 (1993).
12. Müller-Navarra, D. C. Evidence that a highly unsaturated fatty acid limits *Daphnia* growth in nature. *Arch. Hydrobiol.* **132**, 297–307 (1995).
13. Brett, M. T. & Müller-Navarra, D. C. The role of highly unsaturated fatty acids in aquatic food-web processes. *Freshwat. Biol.* **38**, 483–499 (1997).
14. Wood, B. J. B. *Algal Physiology and Biochemistry* (ed. Stewart, W. D. P.) 236–266 (Blackwell Scientific, Oxford, 1974).
15. Cobelas, M. A. & Lechado, J. Z. Lipids in microalgae. A review I. Biochemistry. *Grasas Aceites* **40**, 118–145 (1989).
16. Ahlgren, G., Gustafsson, I. B. & Boberg, M. Fatty acid content and chemical composition of freshwater microalgae. *J. Phycol.* **28**, 37–50 (1992).
17. Arnold, D. E. Ingestion, assimilation, survival, and reproduction by *Daphnia pulex* fed seven species of blue-green algae. *Limnol. Oceanogr.* **16**, 906–920 (1971).
18. Porter, K. G. & Orcutt, J. D. in *Evolution and Ecology of Zooplankton Communities* (ed. Kerfoot, W. C.) 268–281 (University Press, Hanover, 1987).
19. Lampert, W. Studies on the carbon balance of *Daphnia pulex* de Geer as related to environmental conditions. II. The dependence of carbon assimilation on animal size, temperature, food concentration and diet species. *Arch. Hydrobiol. Suppl.* **48**, 310–335 (1977).
20. Vincent, W. F. (ed.) Dominance of bloom forming cyanobacteria (blue-greens). *NZ J. Mar. Freshwat. Res.* **21**, (1987).
21. Lampert, W. Feeding and nutrition in *Daphnia*. (eds Peters, R. H. & De Bernardi, R.) *Mem. Ist. Ital. Idrobiol.* **45**, 143–192 (1987).
22. Lampert, W., Schmitt, R.-D. & Muck, P. Vertical migration of freshwater zooplankton: test of some hypothesis predicting a metabolic advantage. *Bull. Mar. Sci.* **43**, 620–640 (1988).
23. Greenberg, A. E., Trussel, R. R., Clesceri, L. S. & Franson, M. A. H. (eds) *Standard Methods for the Examination of Water and Wastewater* (United Book, Baltimore, 1995).
24. Marker, A. F., Crowther, C. A. & Gunn, R. J. M. Methanol and acetone as solvents for estimating chlorophyll-*a* and pheopigments by spectrophotometry. *Arch. Hydrobiol. Beih. Ergebn. Limnol.* **14**, 52–69 (1980).
25. Kattner, G. & Fricke, H. S. G. Simple gas-liquid chromatographic method for the simultaneous determination of fatty acids and alcohols in wax esters of marine organisms. *J. Chromatogr.* **361**, 263–268 (1986).

Acknowledgements

We thank the members of the Limnological Group at UC Davis for their help, and especially G. Malyj for correcting the English. This work was supported by the US NSF (C.R.G. and M.T.B.).

Correspondence and requests for materials should be addressed to D.C.M.-N. (e-mail: dcnavarra@ucdavis.edu).

The Santa Barbara Basin is a symbiosis oasis

Joan M. Bernhard*, Kurt R. Buck†, Mark A. Farmer‡
& Samuel S. Bowser§

* Department of Environmental Health Sciences, School of Public Health, University of South Carolina, Columbia, South Carolina 29208, USA

† Monterey Bay Aquarium Research Institute, 7700 Sandholdt Road, Moss Landing, California 95039, USA

‡ Department of Cellular Biology, University of Georgia, Athens, Georgia 30602, USA

§ Wadsworth Center, New York State Department of Health, Albany, New York 12201-0509, USA

It is generally agreed that the origin and initial diversification of Eucarya occurred in the late Archaean or Proterozoic Eons when atmospheric oxygen levels were low¹ and the risk of DNA damage due to ultraviolet radiation was high². Because deep water provides refuge against ultraviolet radiation³ and early eukaryotes may have been aerotolerant anaerobes^{4,5}, deep-water dysoxic environments are likely settings for primeval eukaryotic diversification. Fossil evidence shows that deep-sea microbial mats, possibly of sulphur bacteria similar to *Beggiatoa*, existed during that time⁶. Here we report on the eukaryotic community of a modern analogue, the Santa Barbara Basin (California, USA). The *Beggiatoa* mats of these severely dysoxic and sulphidic sediments support a surprisingly abundant protistan and metazoan meiofaunal community, most members of which harbour prokaryotic

Table 1 SBB eukaryotic taxa/morphotypes

Taxa/Morphotype	Method	Ectobionts	Endobionts
Protista			
Foraminifera			
Allogromid sp. 1	DAPI	No	Yes
Allogromid sp. 2	TEM	No	No
<i>Buliminella tenuata</i>	TEM	No	Yes
<i>Chilostomella ovoidea</i>	TEM	No	No
<i>Fursenkoina rotundata</i>	Autofluorescence	No	Yes†
<i>Nonionella stella</i>	TEM	No	Yes‡
Euglenozoan Flagellates			
<i>Calkinsia aureus</i>	DAPI/SEM	Yes	No
<i>Notosolenus ostium</i>	DAPI/SEM	No	No
<i>Postgaardi mariagerensis</i>	DAPI/SEM	Yes	No
<i>Sphenomonas</i> sp.	DAPI	No	No
Euglenoid sp.	DAPI/SEM/TEM	Yes	No
Flagellate sp. 1	DAPI	No	Yes
Flagellate sp. 2	DAPI	Yes	No
Ciliophora			
<i>Frontonia</i> ? sp.	DAPI	No	No
<i>Litonotus duplostriatus</i> ?	DAPI	No	No
<i>Metopus halophila</i>	DAPI	Yes	Yes
<i>Metopus verrucosus</i>	TEM/DAPI	Yes	No
<i>Parablepharisma collare</i> ?	DAPI	Yes	No
<i>Parablepharisma</i> sp.	DAPI	Yes	Yes
Ciliate sp. 1	DAPI	No	Yes
Ciliate sp. 2	DAPI	Yes	No
Ciliate sp. 3	DAPI	Yes	No
Ciliate sp. 4	DAPI	Yes	No
Ciliate sp. 5	DAPI	Yes	No
Ciliate sp. 6	DAPI	Yes	No
Ciliate sp. 7	DAPI	No	Yes
Ciliate sp. 8	DAPI	No	Yes
Ciliate sp. 9	DAPI	Yes	No
Metazoa			
Nematoda			
<i>Desmodora masira</i>	TEM/DAPI	Yes	No
<i>Daptonema</i> sp.	TEM/DAPI	No	No
Gastrotricha			
<i>Urodasys</i> sp.	DAPI/SEM	No	No†
Polychaeta			
<i>Megamerilla</i> ? sp.	TEM/SEM	Yes	No
Gastropoda			
<i>Astryx permodesta</i>	TEM	No	No†§

* Cyanobacteria?

† Preliminary examination; TEM studies pending.

‡ Chloroplasts.

§ Absent in most severely dysoxic samples.

We omit three species of foraminifera (*Bolivina tumida*, *Spiroplectammina earlandi*, *Suggrunda eckis*) because they have not yet been examined with TEM. Also omitted are single or rare occurrences of a given flagellate or ciliate morphotype. Yes, present; No, absent. DAPI, 4,6-diamidino-2-phenylindole; TEM, transmission electron microscopy; SEM, scanning electron microscopy.

symbionts. Many of these taxa are new to science, and both microaerophilic and anaerobic taxa appear to be represented. Compared with nearby aerated sites, the Santa Barbara Basin is a 'symbiosis oasis' offering a new source of organisms for testing symbiosis hypotheses of eukaryogenesis.

The Santa Barbara Basin (SBB), which is a depression between the California mainland and the northern Channel Islands (34° 15' N, 120° 02' W; maximum depth ~600 m, sill depth ~475 m), has dysoxic bottom waters owing to local bathymetry, circulation and a well-developed Oxygen Minimum Zone. The surface sediments of the central SBB are even more oxygen-depleted ($[O_2] \leq 1 \mu M$ within the top 2–3 mm; ref. 7) and sulphide concentrations in the surface centimetre can be more than 0.1 μM (ref. 8). These conditions support a mat of the filamentous sulphide-oxidizing bacterium *Beggiatoa*⁹ but are toxic to mega- and macrofauna except the surface-dwelling gastropod *Astryx permodesta*. Although much is known regarding shallow-water anaerobic and microaerophilic eukaryotes and their symbionts^{10–14}, little is known about comparable deep-water organisms. Furthermore, previous studies of meiofaunal eukaryotes from shallow-water or bathyal oxygen-depleted sediments have not methodically assessed symbioses on a community scale, including all protistan and metazoan groups.

We found that the SBB benthic eukaryotic community includes 36 common taxa or morphotypes (Table 1). Although euglenozoan flagellates have seldom been associated with oxygen-depleted sulphidic marine environments^{10,11,15}, they were numerically the most abundant group (Fig. 1a); non-euglenozoan flagellates were rarely observed. By comparison, foraminifera contributed the highest biovolume—over sixfold higher than that of euglenozoans (Fig. 1d).

Most eukaryotic SBB taxa (24 morphotypes) harboured bacterial ectobionts and/or endobionts (Table 1). The consistent lack of ectobionts on certain taxa indicates that bacteria are not merely using the exteriors of eukaryotes as substrate. Rather, the observed characteristic and highly specific ultrastructural affinities between eukaryotes and prokaryotes, see below, compel us to infer symbiotic relationships. In this context, symbiosis does not necessarily imply mutualism because host–prokaryote metabolic interactions have yet to be determined. The majority of both flagellates and ciliates had bacterial symbionts (Fig. 1b and c) and over 80% of the SBB flagellate biovolume had associated prokaryotes (Fig. 1e). Although the biovolume of ciliates with ectobionts and/or endobionts comprised just over one-third of total ciliate biovolume (Fig. 1f), the majority of ciliate taxa had associated prokaryotes (13 out of 15; Table 1). The relatively low proportion of ciliate biovolume with associated prokaryotes is explained by the fact that the two non-symbiont-bearing ciliate morphotypes were considerably larger, as well as more abundant than those ciliates with bacterial associates.

The SBB samples had meiofaunal abundances and biovolumes that were over an order of magnitude greater than those of more

aerated comparative sites with undetectable sulphide (Fig. 1a, d). The comparative sites, which were located outside the SBB, also had substantially lower proportions of flagellates and ciliates harbouring bacterial ectobionts and/or endobionts (Fig. 1b, c). Only about 20% of the combined flagellate and ciliate biovolume from aerated sites had symbionts (Fig. 1e, f). The average density of eukaryotes in our SBB samples was comparable to those determined for a shallow water *Beggiatoa*-laden site in Denmark¹⁶ but was much higher than protistan and ‘micrometazoan’ densities from aerated intertidal sediments¹⁷ and other aerated bathyal sites¹⁸. The SBB eukaryotic biovolume was about an order of magnitude higher than that of the meiofauna from bathyal cold seeps off central California¹⁵. Our observations represent initial occurrences of many species in deep-sea sediments including the protists *Calkinsia aureus*, *Metopus verrucosus* and *Postgaardia mariagerensis*, as well as the polychaete *Meganerilla*, which were previously recorded from less than 23 m (refs 10–12, 19).

Scanning (SEM) and transmission (TEM) electron microscopy were used to document the structural relationships of prokaryotic–eukaryotic symbioses. Numerous intact, rod-shaped bacteria were seen by TEM throughout the cytoplasm of the foraminifer *Buliminella tenuata* (Fig. 2a; ref. 20). We assert that these bacteria were symbionts because the vast majority (93%) appeared intact and 3% were dividing. Furthermore, if these bacteria represented prey they would be restricted to the youngest chambers, which is the site of digestion in such foraminifera²¹. This adds to the few reports of rhizopod–bacterial symbioses^{20,22,23}, although foraminiferal–algal symbioses are common in photic settings²¹. *Nonionella stella*,

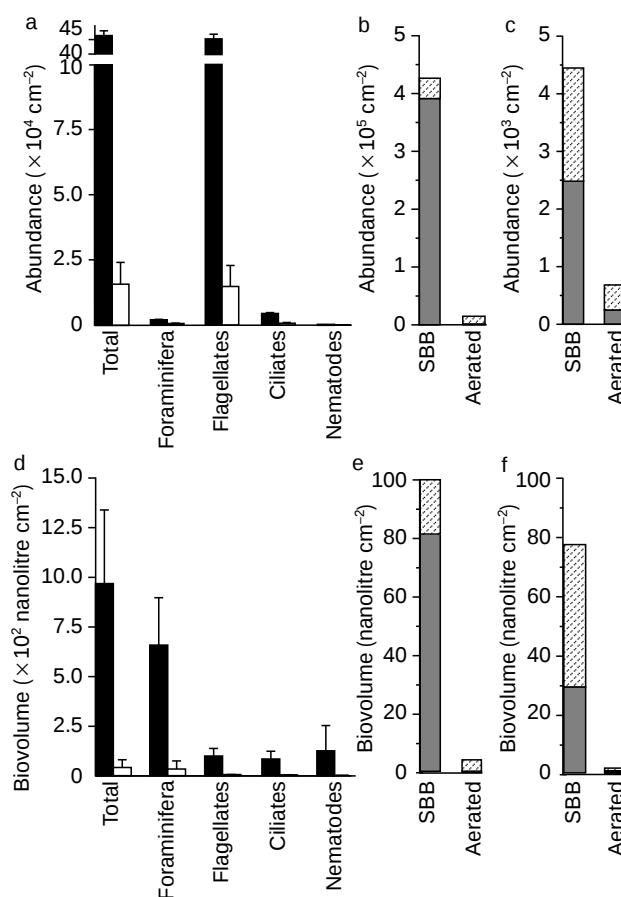


Figure 1 Abundance and biovolume of SBB meiofauna. **a–c**, Abundance of dominant eukaryotic groups (**a**), flagellates (**b**) and ciliates (**c**) in Santa Barbara Basin (SBB) and comparative aerated, sulphide-free samples. **d–f**, Biovolume of dominant eukaryotic groups (**d**), flagellates (**e**), ciliates (**f**). Black bars, SBB; white bars, aerated samples; grey

bars, specimens with ectobionts and/or endobionts; hatched bars, specimens without ectobionts and/or endobionts. [O₂] = 0.2–3.8 μM for SBB and 13.8–25.2 μM for aerated samples. Abundance and biovolume data are not included for gastrotrichs and polychaetes, which had relatively low abundances.

the most abundant benthic foraminifer, possessed intact chloroplasts²⁴ (data not shown), which were numerous enough to impart colouration to the host. Chloroplast sequestration in hosts living far below the euphotic zone lends credence to the notion that other key metabolic roles besides photosynthesis are functioning in these plastids. The rarely documented euglenozoan *Calkinsia aureus* was veiled with bacterial rods that extended off its posterior, appearing as a tail (Fig. 2b). An undescribed Euglenoid species, which is likely to be a representative of a new higher taxon, had ectobionts restricted to its posterior and nestled within pellicular grooves (Fig. 2c). The ciliate *Metopus verrucosus* had 8–12 transverse DAPI-stained bands (inset, Fig. 2d), which TEM revealed to be clusters of bacterial rods oriented normal to its surface (Fig. 2d).

The nematode *Desmodora masira* harboured rod-shaped ectobiotic bacteria in previously undocumented association: the bacteria were aligned, typically end-to-end, under the nematode's annulate cuticle (Fig. 2e). The other SBB nematode, which did not have ectobionts, was rare compared with *D. masira* (comprising ~11% of all nematodes). An undescribed archiannelid polychaete (*Meganerilla?* sp.) possessed a dense layer of rod-shaped bacteria completely vesting its cuticle (Fig. 2f). Because the only other documented cases of polychaetes with bacterial ectosymbionts

occur at hydrothermal vents²⁵, our observations extend this polychaete life strategy to a different extreme environment.

The molecular identification and physiology of the bacteria associated with SBB eukaryotes are currently being studied. We anticipate that diverse metabolic relationships exist because (1) initial ultrastructural findings point toward sulphide-oxidizing bacteria on at least some taxa; (2) many previously described ciliate ectobionts are sulphate reducers¹³; and (3) all well-studied *Metopus* species are known to harbour methanogens¹². Because methanogens may have played a crucial role in the evolution of eukaryotes^{4,5}, *Metopus* may be important to study from an evolutionary perspective. The SBB provides the first deep-water source of abundant *Metopus* species.

The large number and variety of prokaryotic–eukaryotic structural associations in the SBB establish it as a new model community for studying chemoautotrophic symbioses. Of particular interest are the parallels to ancient symbioses that might have led to early eukaryotic diversification by lateral gene transfer. Much work needs to be done on the eubacterial and archaeal communities of the SBB sediments²⁶, especially those that might be extant homologues to presumptive progenotes. A related possibility is that laminated SBB deposits, which are ~200 m in thickness²⁷, might provide source fossils for comparisons with those of Proterozoic strata. Although mineralized microfossils of the SBB have been extensively studied²⁷, soft-bodied preservation in the basin has not yet been documented.

Just as bacterial–eukaryotic symbiotic communities could have been widespread in the past, they are probably more widespread in the modern ocean than is presently recognized, because environments supporting chemoautotrophic bacterial mats, such as silled basins, hydrocarbon seeps, hydrothermal vents and Oxygen Minimum Zones are common in today's oceans. It is likely, therefore, that bacterial–eukaryotic symbiotic communities like those found in the SBB have an underestimated and significant impact on oceanographic processes such as nutrient cycling. Anthropogenic activities are expanding oxygen-depleted habitats²⁸, and concomitantly such communities are expected to become even more widespread in the future. Detailed knowledge about the inhabitants and processes occurring in these communities will be required to effect bioremediation.

Methods

Samples were collected with either a Soutar boxcorer or Multicorer. Dissolved oxygen gas was determined using microwinkler analysis²⁹ of seawater collected in a Niskin bottle attached to the corer frame. Samples denoted 'SBB' were collected from an area of laminated, *Beggiatoa*-laden sediments delimited roughly by 34° 12' N, 34° 17' N, 119° 57' W and 120° 03' W. Comparative 'Aerated' sites were located off Palos Verde (~33° 46.7' N, 118° 32.4' W) and Point Conception (~34° 18.0' N, 120° 41.3' W). SBB abundance and biovolume data are the mean of 5 replicates collected in 2/98, 9/98 and 2/99; Aerated data represent 3 samples, one collected in each of the same months.

Abundance and biovolume data were obtained from 60-cm³ syringe subcores from which the surface centimetre was removed and fixed in chilled 3% glutaraldehyde/0.1 M Na-cacodylate buffer pH 7.2. For quantitative analyses, organisms were isolated from sample splits using a Percoll gradient technique³⁰, concentrated on blackened 3-µm Nucleopore filters, stained with the DNA-intercalating dye 4,6-diamidino-2-phenylindole (DAPI), and observed using either a Nikon Diaphot equipped with UV-1A filter cube and 20x Fluor objective lens (0.75 NA) or an epifluorescence-equipped Zeiss Axioskop. At the appropriate magnification, cells were counted, dimensions noted and assessed for presence/absence of ectobionts and/or endobionts. To distinguish ciliate and flagellate morphotypes, organisms were observed with differential interference contrast (DIC) optics. Our morphogroup assignments were conservative because ciliates are morphologically plastic³¹. Because organism and sample size must be compatible to ensure accurate abundance determinations, the densities of larger taxa may be underestimates. Assessment of associated prokaryotes in metazoans and foraminifera was accomplished using TEM and/or SEM (see below) because the DAPI signal from metazoan nuclei and the opacity of foraminiferal tests mask bacterial DAPI signals.

For ultrastructural studies, surface sediment was placed in chilled glutaraldehyde/cacodylate buffer within 10 minutes of collection. In the laboratory, sediments were sieved over a 63-µm screen using chilled buffer. Specimens were removed from the coarse fraction using a stereomicroscope. After osmication, decalcification (only for foraminifera), ethanol dehydration and embedding in Epon-Araldite, 70-nm sections were examined with a Philips 301 TEM. Before fixation of *Metopus verrucosus*, sediments were incubated in Cell-Tracker Green CMFDA (Molecular Probes, Eugene, OR, USA) for 4 hours. The specimen was isolated after observation with an epifluorescence-equipped

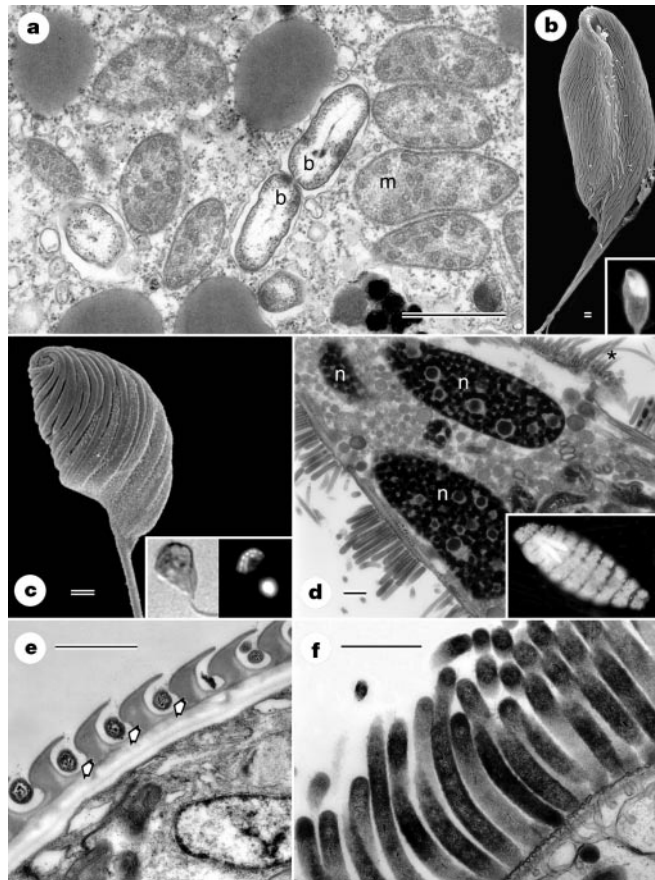


Figure 2 Micrographs of SBB meiofauna with symbionts. **a**, TEM showing bacterium dividing in *Buliminella tenuata*. m, mitochondria; b, bacterium. **b**, SEM showing bacteria vesting *Calkinsia aureus*. Inset, DAPI-stained specimen, showing bacteria and flagellate nucleus. **c**, SEM showing bacteria nestled in pellicle of Euglenoid species. Inset, DIC- and DAPI-stained specimen, showing bacteria and flagellate nucleus. **d**, HVEM showing bacterial aggregations on the surface of *Metopus verrucosus*. n, macronuclei; star, cilia. Inset, DAPI-stained *M. verrucosus*, showing bacterial bands and ciliate nuclei. **e**, TEM of cross-section through *Desmodora masira* cuticle showing bacteria (arrows) under annuli. **f**, TEM showing bacteria attached to exterior of *Meganerilla?* sp. All scale bars, 1 µm. DIC, differential interference contrast; HVEM, high-voltage electron micrograph.

Nikon Optiphot. After placement on a cover slip, it was overlain with agar, and subsequently osmicated, dehydrated and embedded as described. Sections of 1 μm in thickness were examined with a High Voltage Electron Microscope (HVEM) at 1 MeV. SEM of flagellates and ciliates involved fixation of ~ 1 ml sediment in 1% OsO_4 vapour for 1 hour, subsequent concentration onto 3- μm filters, and dehydration with ethanol. Filters were then critical-point dried, coated with Au and examined with a LEO 982.

- Knoll, A. H. & Holland, H. D. in *Effects of Past Global Change on Life* 21–33 (National Research Council, Washington, DC, 1995).
- Kasting, J. F. Earth's early atmosphere. *Science* **259**, 920–926 (1993).
- Cockell, C. S. Biological effects of high ultraviolet radiation on early Earth—a theoretical evaluation. *J. Theor. Biol.* **193**, 717–729 (1998).
- Martin, W. & Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41 (1998).
- Moreira, D. & López-García, P. Symbiosis between methanogenic Archaea and δ -Proteobacteria as the origin of eukaryotes: The syntrophic hypothesis. *J. Mol. Evol.* **47**, 517–530 (1998).
- Simonson, B. M. & Carney, K. E. Roll-up structures: evidence of *in situ* microbial mats in Late Archean deep shelf environments. *Paleo* **14**, 13–24 (1999).
- Reimers, C. E. *et al.* Porewater pH and authigenic phases formed in the uppermost sediments of the Santa Barbara Basin. *Geochim. Cosmochim. Acta* **60**, 4037–4057 (1996).
- Kuwabara, J. S. *et al.* Dissolved sulfide distributions in the water column and sediment pore waters of the Santa Barbara Basin. *Geochim. Cosmochim. Acta* **63**, 2199–2209 (1999).
- Soutar, A. & Crill, P. A. Sedimentation and climatic patterns in the Santa Barbara Basin during the 19th and 20th centuries. *Geol. Soc. Am. Bull.* **88**, 1161–1172 (1977).
- Fenchel, T. *et al.* Microbial diversity and activity in a Danish fjord with anoxic deep water. *Ophelia* **43**, 45–100 (1995).
- Simpson, A. G. B. *et al.* The ultrastructure and systematic position of the euglenozoon *Postgaardi mariagerensis*, Fenchel *et al.* *Arch. Protistenkd.* **147**, 213–225 (1996/1997).
- Esteban, G., Fenchel, T. & Finlay, B. Diversity of free-living morphospecies in the ciliate genus *Metopus*. *Arch. Protistenkd.* **146**, 137–164 (1995).
- Fenchel, T. & Finlay, B. J. (eds) *Ecology and Evolution in Anoxic Worlds* (Oxford Univ. Press, Oxford, 1995).
- Ott, J. A., Novak, R., Schiemer, F., Hentschel, U., Nebelsick, M. & Polz, M. Tackling the sulfide gradient—a novel strategy involving marine nematodes and chemoautotrophic ectosymbionts. *PSZNI Mar. Ecol.* **12**, 261–279 (1991).
- Buck, K. R. & Barry, J. P. Monterey Bay cold seep infauna: quantitative comparison of bacterial mat meiofauna with non-seep control sites. *Cahiers de Biologie Marine* **39**, 333–335 (1998).
- Bernard, C. & Fenchel, T. Mats of colourless sulphur bacteria: II. Structure, composition of biota and successional patterns. *Mar. Ecol. (Prog. Ser.)* **128**, 171–179 (1995).
- Epstein, S. Simultaneous enumeration of protozoa and micrometazoa from marine sandy sediments. *Aquat. Microb. Ecol.* **9**, 219–227 (1995).
- Alongi, D. M. The distribution and composition of deep-sea microbenthos in a bathyal region of the western Coral Sea. *Deep-Sea Res.* **34**, 1245–1254 (1987).
- Saphonov, M. V. & Tzvetlin, A. B. Nerillidae (Annelida: Polychaeta) from the White Sea, with description of a new species of *Micronerilla* Jouin. *Ophelia* **47**, 215–226 (1997).
- Bernhard, J. M. Microaerophilic and facultative anaerobic benthic foraminifera: a review of experimental and ultrastructural evidence. *Rev. Paleobiol.* **15**, 261–275 (1996).
- Lee, J. J. & Anderson, O. R. (eds) *Biology of Foraminifera* (Academic, London, 1991).
- Anderson, O. R. & Matsuoka, A. Endocyttoplasmic microalgae and bacteroids within the central capsule of the radiolarian *Dictyocoryne truncatula*. *Symbiosis* **12**, 237–247 (1992).
- Richardson, S. L. & Rützel, K. Bacterial endosymbionts in the agglutinating foraminiferan *Spiculidendron coralliculum* Rützel and Richardson, 1996. *Symbiosis* **26**, 299–312 (1999).
- Bernhard, J. M. & Bowser, S. S. Benthic foraminifera of dysoxic sediments: chloroplast sequestration and functional morphology. *Earth Sci. Rev.* **46**, 149–165 (1999).
- Desbruyères, D. *et al.* Biology and ecology of the “Pompeii worm” (*Alvinella pompejana* Desbruyères and Laubier), a normal dweller of an extreme deep-sea environment: a synthesis of current knowledge and recent developments. *Deep-Sea Res. II* **45**, 383–422 (1998).
- Stolz, J. F., Chang, S.-B. R. & Kirschvink, J. L. Magnetotactic bacteria and single-domain magnetite in hemipelagic sediments. *Nature* **321**, 849–851 (1986).
- Kennett, J. P., Baldauf, J. G. & Lyle, M. (eds) *Proceedings of the Ocean Drilling Program, Scientific Results Vol. 146 Part 2* (Ocean Drilling Program, College Station, TX, 1995).
- Diaz, R. J. & Rosenberg, R. Marine benthic hypoxia: a review of its ecological effects and the behavioural responses of benthic macrofauna. *Oceanogr. Mar. Biol. Annu. Rev.* **33**, 245–303 (1995).
- Broenkow, W. W. & Cline, J. D. Colorimetric determination of dissolved oxygen at low concentrations. *Limnol. Oceanogr.* **14**, 450–454 (1969).
- Starink, M. *et al.* Quantitative centrifugation to extract benthic protozoa from freshwater sediments. *Appl. Environ. Microbiol.* **60**, 167–173 (1994).

Acknowledgements

We thank the captain, crew and scientific parties of the RV *Robert Gordon Sproul*; C. Bernard, G. Esteban, T. Fenchel and A. Simpson for flagellate and ciliate identifications; W. Fowle, H. Ghiradella, L. Harris, C. Jouin-Toulmond, M. Müller, A. Todaro, A. Vanreusel and M. Vincx for assistance with metazoan identifications; E. Braun-Howland and A. Marcario for valuable discussions; A. van Geen for two collecting opportunities; and T. Fenchel for comments on the manuscript. The Wadsworth Center's Electron Microscopy Core and NIH Biotechnological Resource Grant supporting the Biological Microscopy and Image Reconstruction Resource are acknowledged. This work was supported by USC's Career Development Professorship Program, the Packard Foundation and grants from NSF (PEET Systematics to M.A.F., Polar Programs to S.S.B., Ocean Sciences to J.M.B.).

Correspondence and requests for materials should be addressed to J.M.B. (e-mail: jmbernha@sph.sc.edu).

Mortality of sea lions along the central California coast linked to a toxic diatom bloom

Christopher A. Scholin[†], Frances Gulland[†], Gregory J. Doucette[‡], Scott Benson[§], Mark Busman[‡], Francisco P. Chavez^{*}, Joe Cordaro^{||}, Robert DeLong[¶], Andrew De Vogelaere[#], James Harvey[§], Martin Haulena[†], Kathi Lefebvre[‡], Tom Lipscomb^{**}, Susan Loscutoff^{††}, Linda J. Lowenstine^{‡‡}, Roman Marin III^{*}, Peter E. Miller^{*}, William A. McLellan^{§§}, Peter D. R. Moeller[‡], Christine L. Powell[‡], Teri Rowles^{|||}, Paul Silvagni^{‡‡}, Mary Silver[☆], Terry Spraker^{¶¶}, Vera Trainer^{##} & Frances M. Van Dolah[‡]

^{*} Monterey Bay Aquarium Research Institute, 770 Sandholdt Rd, Moss Landing, California 95039, USA

[†] The Marine Mammal Center, Marin Headlands, Sausalito, California 94965, USA

[‡] Marine Biotoxins Program, NOAA/National Ocean Service, 219 Fort Johnson Rd, Charleston, South Carolina 29412, USA

[§] Moss Landing Marine Laboratories, PO Box 450, Moss Landing, California 95039, USA

^{||} National Marine Fisheries Service, 501 West Ocean Blvd., Suite 4,200, Long Beach, California 90802, USA

[¶] National Marine Mammal Laboratory, 7,600 Sandpoint Way NE, Seattle, Washington 98115, USA

[#] Monterey Bay National Marine Sanctuary, 299 Foam St., Monterey, California 93940, USA

[☆] Institute of Marine Science, University of California at Santa Cruz, Santa Cruz, California 95064, USA

^{**} Armed Forces Institute of Pathology, Veterinary Pathology, Washington, District of Columbia 20306, USA

^{††} California Department of Health Services Food and Drug Branch, PO Box 942732 MS 357, Sacramento, California 94234-7320, USA

^{‡‡} Department of Pathology, Microbiology and Immunology, University of California at Davis, Davis, California 95616, USA

^{§§} Biological Sciences, University of North Carolina at Willmington, 601 S. College Rd., Willmington, North Carolina 28403, USA

^{|||} National Marine Fisheries Service, 1315 East West Highway, Silver Spring, Maryland 20910, USA

^{¶¶} College of Veterinary Medicine, Colorado State University, Fort Collins, Colorado 80523, USA

^{##} NOAA/NMFS/ECDC, 2,725 Montlake Blvd., Seattle, Washington 98112, USA

Over 400 California sea lions (*Zalophus californianus*) died and many others displayed signs of neurological dysfunction along the central California coast during May and June 1998. A bloom of *Pseudo-nitzschia australis* (diatom) was observed in the Monterey Bay region during the same period. This bloom was associated with production of domoic acid (DA), a neurotoxin¹ that was also detected in planktivorous fish, including the northern anchovy (*Engraulis mordax*), and in sea lion body fluids. These and other concurrent observations demonstrate the trophic transfer of DA resulting in marine mammal mortality. In contrast to fish, blue mussels (*Mytilus edulis*) collected during the DA outbreak contained no DA or only trace amounts. Such findings reveal that monitoring of mussel toxicity alone does not necessarily provide adequate warning of DA entering the food web at levels sufficient to harm marine wildlife and perhaps humans.

Harmful algal blooms (HABs) pose serious public health threats and their economic impact can total hundreds of millions of dollars annually^{2,3}. HABs are also associated with mortality of wildlife, including birds and marine mammals^{4–7}. Establishing an unambiguous connection between HABs and marine mammal mortality is difficult. Lack of background data on the sensitivity of these animals to algal toxins, as well as challenges associated with ascribing a definitive cause of death, can make proof of HAB-related marine

mammal mortality elusive⁸. Logistical problems associated with detecting harmful algae and their toxins over large spatial and temporal scales also hamper efforts to map blooms and provide a context within which to evaluate concurrent mortality events. Here, *Pseudo-nitzschia* species-specific DNA probes^{9–11} and a DA receptor binding assay¹² were used to detect a toxic algal bloom. We also employed liquid chromatography–tandem mass spectroscopy (LC–MS/MS) to confirm the presence of DA in plankton, anchovy, and in sea lions that died during the bloom, and histopathology to show brain lesions characteristic of DA poisoning.

The diatoms observed in Monterey Bay in the early spring of 1998 consisted primarily of *Chaetoceros* spp. During this period, species

of *Pseudo-nitzschia* were largely absent and plankton samples did not contain detectable DA activity (see Fig. 1a–c, 11 March–20 April 1998). During the first half of May there was a sudden and sharp rise in the abundance of *P. australis* and in DA activity (Fig. 1a, b). The maximum recorded density of *P. australis* was $\sim 1.3 \times 10^5$ cells l^{-1} . At the height of the bloom, plankton were concentrated in a narrow band along the shore (Fig. 2a) and may have been entrained within, or responded to, a water mass enriched with silicate, a chemical signature indicative of terrestrial freshwater runoff¹³. This pattern is evident in nutrient levels (Fig. 1d) observed at the Santa Cruz sampling station during the initiation phase of the bloom¹⁴. Potential links between *Pseudo-nitzschia* blooms and

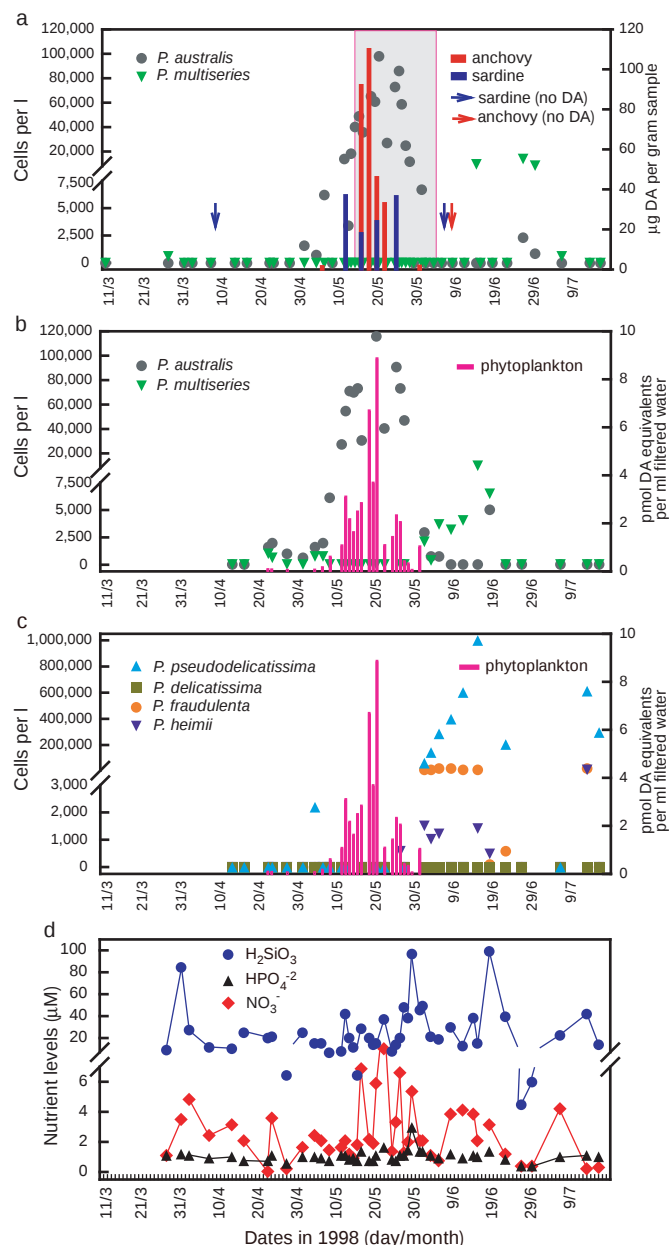


Figure 1 Abundance of *Pseudo-nitzschia* species, DA activity in plankton and fish, and nutrient concentrations. Data taken at Santa Cruz wharf from 11 March–16 July 1998. **a**, *P. australis* and *P. multiseriata* detected by sandwich hybridization^{10,11}. Bars, µg DA per gram sardine and anchovy determined by HPLC–UV (ref. 1; S.L. and G. Langlois, personal communication); arrows, sardine and anchovy collected, DA not detected; shaded area period of sea lion mortality in Monterey Bay. **b**, *Pseudo-nitzschia* species detected by whole-cell hybridization⁹. Bars, receptor assay-based DA activity¹² in plankton. **c**, *Pseudo-nitzschia* species detected by whole-cell hybridization⁹; DA as in **b**. **d**, Concentrations of silicate (H_2SiO_3), nitrate (NO_3^-) and phosphate (HPO_4^{2-}) (ref. 23).

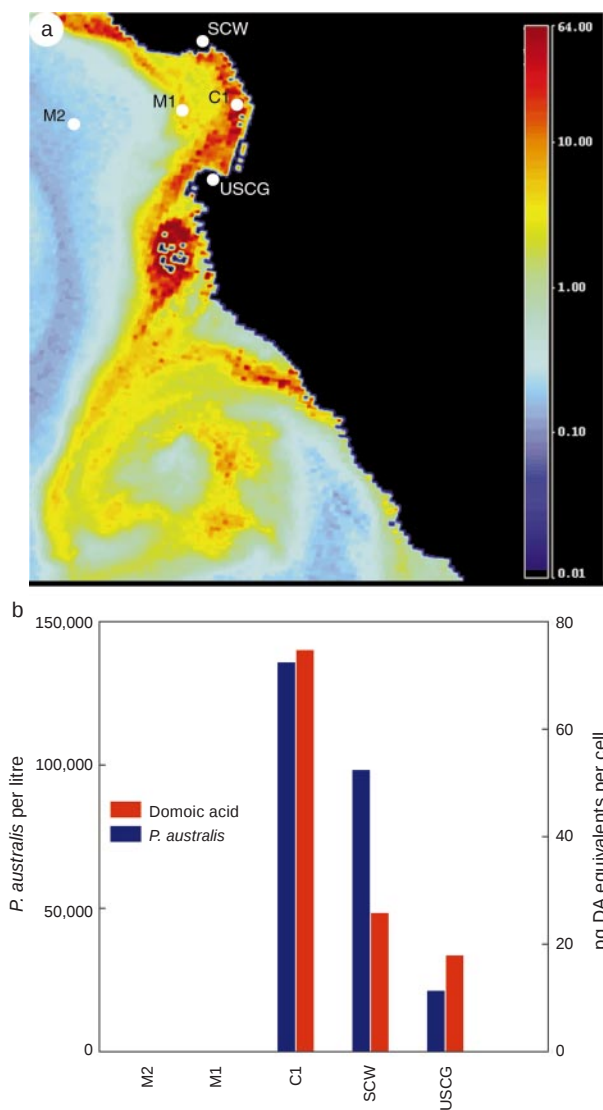


Figure 2 Satellite image of Monterey Bay, California, *P. australis* cell density and DA activity. **a**, 15 May 1998, estimate of chlorophyll *a* inferred from SeaWiFS image of the central California coast produced using SeaDAS with NASA standard algorithms²⁶. Scale is $\mu g\ l^{-1}$ provided for qualitative purposes only as calibration is on-going (<http://www.mbari.org/oasis/index.html>). Shown are the Santa Cruz wharf (SCW) and the United States Coast Guard pier (USCG), C1, M1 and M2, ship and mooring-based sampling sites. **b**, Distribution of *P. australis*^{10,11} and per cell estimates of DA activity¹² (Table 1) on 20 May 1998, at locations shown in **a**.

Table 1 Results of the domoic acid (DA) receptor binding assay for phytoplankton, anchovy and sea lion matrices

Sample identification*	Sample matrix	DA equivalents†	LC-MS/MS‡
SCW 9 May 1998	Phytoplankton	31.18 pg cell ⁻¹	positive
SCW 12 May 1998	Phytoplankton	16.10 pg cell ⁻¹	
SCW 14 May 1998	Phytoplankton	7.22 pg cell ⁻¹	
SCW 18 May 1998	Phytoplankton	20.83 pg cell ⁻¹	
C1 20 May 1998	Phytoplankton	74.59 pg cell ⁻¹	
SCW 25 May 1998	Phytoplankton	8.43 pg cell ⁻¹	positive
Anchovy 22 May 1998a	Anchovy gut	71.30 µg g ⁻¹	
Anchovy 22 May 1998b	Anchovy gut	69.67 µg g ⁻¹	
Anchovy 4 June 1998a	Anchovy gut	2.53 µg g ⁻¹	
Anchovy 4 June 1998b	Anchovy gut	0.27 µg g ⁻¹	nd
CSL 3734 22 May 1998	Sea lion serum	0.20 µg ml ⁻¹	
CSL 3724 21 May 1998	Sea lion serum	0.17 µg ml ⁻¹	positive
CSL 3707 20 May 1998	Sea lion urine	3.72 µg ml ⁻¹	
CSL 3794 25 May 1998	Sea lion urine	1.40 µg ml ⁻¹	
CSL 3749 22 May 1998	Sea lion urine	0.72 µg ml ⁻¹	
CSL 3741 23 May 1998	Sea lion urine	0.03 µg ml ⁻¹	
CSL 3800 26 May 1998	Sea lion urine	0.49 µg ml ⁻¹	
CSL 3726 21 May 1998	Sea lion urine	0.12 µg ml ⁻¹	
CSL 3783 23 May 1998	Sea lion faeces	182.01 µg ml ⁻¹	positive
CSL 3734 22 May 1998	Sea lion faeces	1.31 µg ml ⁻¹	
CSL 3758 23 May 1998	Sea lion faeces	96.08 µg ml ⁻¹	

Note the range of DA equivalent concentrations found in samples testing positive for that toxin. One sample representing each sample type was also screened for DA using LC-MS/MS. The amount of DA activity found in all phytoplankton samples is plotted in Fig. 1b; results listed here are for samples dominated by *P. australis* only. A complete list of sea lion samples examined with corresponding results of DA analyses is given in ref. 17.

* SCW, phytoplankton samples from Santa Cruz Wharf; C1, phytoplankton samples from station C1 (Fig. 2a); anchovy, anchovies harvested from Monterey Bay, stomachs dissected; CSL, California sea lion, The Marine Mammal Center identification number¹⁷.

† Concentration of DA equivalent to a certified reference standard¹²; concentrations for phytoplankton were calculated based on the number of *P. australis* cells in a sample as estimated by whole cell¹⁰ and sandwich¹¹ hybridization.

‡ Positive, confirmation of the presence of DA without quantification; nd, not detected; unexposed control samples of each matrix type were shown to be negative.

freshwater inputs have been noted elsewhere¹⁵. Most estimates of cellular toxin levels ranged between ~7–32 pg DA equivalents per *P. australis* cell, similar to that reported previously¹⁶. However, one sample was estimated to have as much as ~75 pg DA equivalents per *P. australis* cell (Table 1, Fig. 2b). A similarly high per cell value of DA associated with a natural population of *P. australis* in California coastal waters was also recorded in 1998 (V.T. *et al.*, unpublished observation). As the *P. australis* bloom declined, a variety of other *Pseudo-nitzschia* species flourished. *Pseudo-nitzschia pseudodelicatissima* was especially abundant, reaching a maximum density of ~10⁶ cells l⁻¹ (Fig. 1c, mid-June 1998). During this transition in species composition, DA activity declined precipitously and became undetectable (Fig. 1c).

Blue mussels (*Mytilus edulus*) collected from Santa Cruz during the *P. australis* bloom and tested using the currently accepted high-performance liquid chromatography-ultraviolet (HPLC-UV) regulatory method¹, contained no DA or only trace quantities (0–2 µg g⁻¹; S.L. and G. Langlois, personal communication) and never exceeded levels considered hazardous for human consumption (≥20 µg g⁻¹)¹. In sharp contrast to mussels, anchovies and sardines caught from Monterey Bay during the *P. australis* bloom and analysed by HPLC-UV contained much more toxin (~30–110 µg g⁻¹; S.L. and G. Langlois, personal communication). The rise and fall of DA in these fish corresponded to the appearance and disappearance of *P. australis* (Fig. 1a). Stomachs of anchovies collected at the peak of the bloom (~22 May 1998) contained high levels of DA activity (Table 1) as well as very high numbers of *P. australis* frustules, but not frustules of other *Pseudo-nitzschia* species (data not shown). Stomach contents of anchovies caught in Monterey Bay also reflected the marked decline of *P. australis* and sudden rise of *P. pseudodelicatissima* (data not shown), along with the corresponding drop in DA activity (4 June 1998; Table 1).

The number of bird and marine mammal carcasses found along the beaches of Monterey Bay increased markedly during the *P. australis* bloom¹⁷. The most frequently encountered marine

mammal was the California sea lion, and live animals of this species were reported as suffering neurological dysfunction. Although over 400 sea lion carcasses were observed between 18 May and 19 June 1998, only 70 live sea lions and one northern fur seal (*Callorhinus ursinus*) were examined clinically and histopathologically at The Marine Mammal Center (TMMC). Of those 70 animals, 28 were found along the beaches of Monterey Bay; of those 28 animals, 20 were collected between 24 and 30 May (ref. 17), a time when the *P. australis* cell density and DA activity had peaked. Reports of live, sick sea lions in Monterey Bay decreased sharply from the end of May onwards, and ceased altogether after 5 June, coinciding with the observed decline in *P. australis* cell density and DA activity (Fig. 1a, b).

Of the 70 animals cared for by TMMC, 48 died despite treatment. Those that recovered appeared clinically normal. All affected animals displayed similar neurological symptoms including seizures, head weaving, ataxia, depression and abnormal scratching¹⁷, the latter being reminiscent of the scratching behaviour documented in mice and pelicans poisoned with DA¹. The animals were also in good body condition and had haematological and serum biochemical parameters within normal limits¹⁷. Histological examination of tissues from all 48 sea lions that died revealed unique lesions in the brain and heart. Brain lesions, characterized by zonal vacuolation of the hippocampal neuropile involving several architectural strata, were most severe in the anterior ventral hippocampus (Fig. 3). Such lesions are similar to those observed in mice, macaques and humans exposed to DA^{18–20}.

Sea lion urine, serum or faeces from 48 animals were tested for DA by receptor-binding assay, with selected samples analysed by LC-MS/MS (Table 1). Receptor-based toxin activity was detected in each sample type, but not from all animals examined¹⁷. Results of LC-MS/MS analyses established the presence of DA in sea lion urine and faeces, and represent the first unequivocal localization of this toxin in mammalian body fluids in connection with a DA-implicated mortality event. Faecal samples from two sea lions in which DA was detected also contained *P. australis* frustules (K.L., unpublished observation), providing an additional link between sickened animals and the DA-producing diatom.

Unambiguous confirmation of DA in samples of *P. australis*-dominated phytoplankton, northern anchovy and sea lion body fluids provides evidence for the trophic transfer of this toxin from its algal source to a marine mammal via a fish vector. This conclusion is strengthened further by the temporal association of toxic *P. australis* with this mortality event (Fig. 1a and b), the fact that anchovies are a well known prey of sea lions²¹, the finding of *P. australis* frustules and anchovy vertebrae in some faecal samples of sickened sea lions¹⁷, detection of DA in body fluids of some sea lions suffering neurological dysfunction, and brain lesions in dead sea lions consistent with exposure to harmful levels of DA (Fig. 3). The total number of sea lions that may have suffered from DA poisoning is not known, but is certainly much greater than the 70 individuals brought to TMMC. Much of the coast affected by the bloom is inaccessible or infrequently visited, counts of dead animals were restricted to specific beaches, no comprehensive land or sea-based surveys for affected animals were undertaken, and only live stranded individuals reported to and retrieved by TMMC were examined. Consequently, the full impact of this bloom on the sea lion population is difficult to gauge.

Although this report provides the first conclusive evidence linking sea lion deaths to a documented HAB, events similar to those described herein appear to have occurred previously. For example, Ocha *et al.*²² have described dolphin and sea lion mortality associated with *Pseudo-nitzschia* blooms in Mexico, but the cause of death was not established conclusively. Along the California coast, adult sea lions and northern fur seals suffering from neurological dysfunction similar to that documented here were also noted in 1978, 1986, 1988 and 1992 (ref. 17). Animals that died in 1992

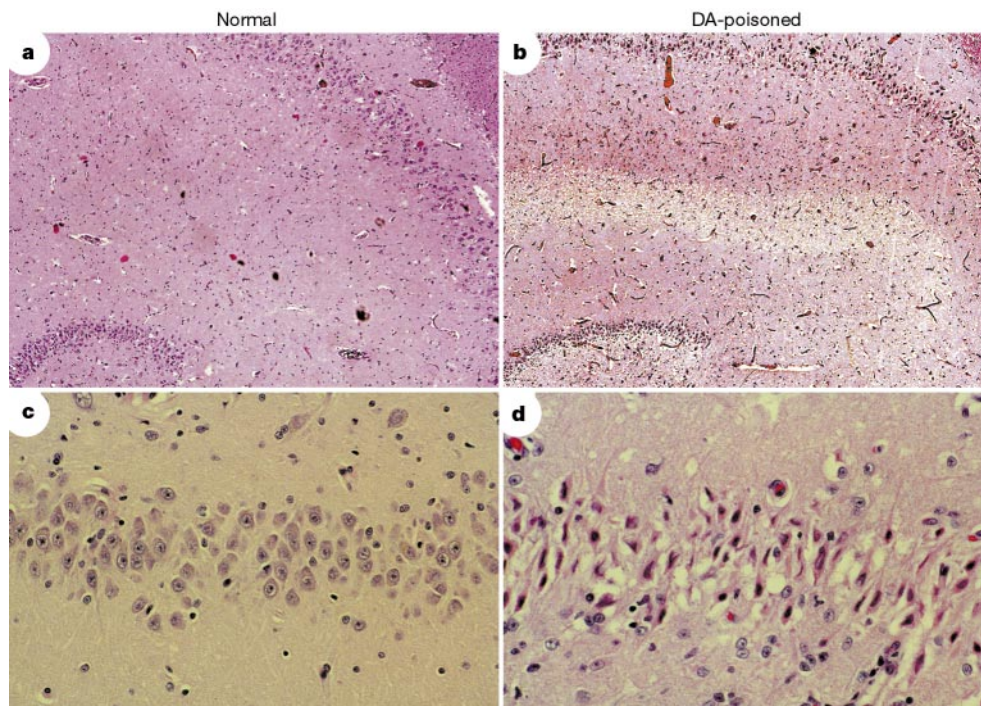


Figure 3 Histological sections of the anterior ventral hippocampus from normal and DA-poisoned California sea lions. Sections prepared with haematoxylin and eosin²⁷.

a, Anterior ventral hippocampus from a normal animal (original magnification 13.2×).

b, Section as in **a**, from a DA-poisoned animal, showing pronounced laminar pallor due to

vacuolation involving several strata (radiatum, lauculosum and moleculare of both the hippocampus and dentate gyrus). **c**, Close-up of dentate gyrus from a normal animal (original magnification 100×). **d**, Section as in **c**, from a DA-poisoned animal showing acute neuronal necrosis and vacuolation.

displayed brain lesions characteristic of DA poisoning, although the connection to this toxin was not established at that time. Toxic blooms of *Pseudo-nitzschia* can be expected in the future along the shores of California and elsewhere. The rapid and dramatic nature of the 1998 mortality event suggests that even a short 'pulse' of DA in the food web, lasting only days or weeks in a localized area, may be sufficient to kill or sicken marine wildlife. □

Methods

Plankton analysis and nutrient determinations

Whole-water (unconcentrated) plankton samples were collected from the surface at locations shown in Fig. 2a and were subjected to DNA probe-based tests using whole-cell⁹ and sandwich^{10,11} hybridization techniques to identify and quantify a variety of toxic and nontoxic *Pseudo-nitzschia* species. Reagents and instrumentation needed for the sandwich hybridization assay are available commercially (Saigene). Aliquots of the water samples were also subjected to DA¹² and nutrient²³ analyses. Data shown here are for the SCW station only.

Domoic acid analysis

Particulate material in whole-water field samples (150–750 ml) was collected onto 25 mm GF/F filters (Whatman), then frozen at –70 °C. Filters were extracted in 10% aqueous methanol and tested using a DA receptor binding assay¹². The following modifications were implemented: concentrations and volumes of reagents used in the glutamate decarboxylase (GAD) digestion to eliminate ambient glutamate from extracts were adjusted to provide a two-fold sample dilution; assays were performed in 96-well MultiScreen plates (Millipore), after which 25 µl of Optiphase liquid scintillant were added to each well and plates counted directly on a microplate scintillation counter (Microbeta 1450, Wallac, Inc.). The DACS-1B certified DA reference standard (National Research Council, Halifax) was used to calibrate the assay and determine equivalent DA activity in a sample. Values are expressed as pmol DA equivalents per ml whole water filtered and indicate DA activity in the ≥0.7 µm fraction normalized to volume filtered.

Detection of DA activity in sea lion faeces and anchovy samples by receptor binding assay required an initial extraction with 50% aqueous methanol²⁴. Faecal extracts (only) were further cleaned using a strong anion exchange (SAX) solid-phase extraction protocol²⁵. Anchovy and cleaned faecal extracts, as well as native sea lion serum and urine samples, were subjected to a GAD digest and run on the receptor assay as described above. The sample of each type showing the highest DA concentration was then analysed by LC-MS/MS to confirm the presence of the toxin. The same extracts or native samples (without GAD digestion) were fractionated on a C18 column using a gradient of 1–95% methanol in 0.1% TFA. A PE SCIEX API-III triple quadrupole mass spectrometer was employed, with the ionspray source operated in positive ion mode, using compressed air as the

nebulization gas. The first quadrupole was used to pass only ions of nominal 312 *m/z*. The second quadrupole was used to facilitate a substantial amount of collisionally induced dissociation, while the third quadrupole was operated in multiple ion monitoring mode where fragment ions of 161 and 266 *m/z*, as well as residual parent ions (312 *m/z*), were allowed to pass to the ion detector.

Received 21 June; accepted 1 November 1999.

- Wright, J. L. C. & Quilliam, M. A. in *Manual on Harmful Marine Microalgae* (eds Hallegraeff, G. M., Anderson, D. M. & Cembella, A. D.) 113–133 (Intergovernmental Oceanographic Commission Manuals and Guides No. 33, UNESCO, Paris, 1995).
- Boesch, D. F. *et al.* *Harmful Algal Blooms in Coastal Waters: Options for Prevention, Control and Mitigation* (NOAA Coastal Ocean Program Decision Analysis Series No. 10 NOAA Coastal Office, Silver Spring, Maryland, 1997).
- Anderson, D. M., Galloway, S. B. & Joseph, J. D. *Marine Biotoxins and Harmful Algae: A National Plan* (Woods Hole Oceanographic Institution, Technical Report 93-02, Woods Hole, Massachusetts, 1993).
- Work, T. M. *et al.* Epidemiology of domoic acid poisoning in brown pelicans (*Pelicanus occidentalis*) and Brandt's cormorants (*Phalacrocorax penicillatus*) in California. *J. Zoo Wild. Med.* **24**, 54–62 (1993).
- Sierra-Beltrán, A. P., Palafox-Urbe, M., Grajales-Montiel, J., Cruz-Villacorta, A. & Ochoa, J. L. Sea bird mortality at Cabo San Lucas, Mexico: evidence that toxic diatom blooms are spreading. *Toxicol.* **35**, 447–453 (1997).
- Bossart, G. D., Baden, D. G., Ewing, R. Y., Roberts, B. & Wright, S. Brevetoxicosis in manatees (*Trichechus manatus latirostris*) from the 1996 epizootic: gross, histologic, and immunohistologic features. *Toxicol. Pathol.* **26**, 276–282 (1998).
- Geraci, J. J. *et al.* Humpback whales (*Megaptera novaeangliae*) fatally poisoned by dinoflagellate toxin. *Can. J. Fish. Aquat. Sci.* **46**, 1895–1898 (1989).
- Harwood, J. What killed the monk seals? *Nature* **393**, 17–18 (1998).
- Miller, P. E. & Scholin, C. A. Identification and enumeration of cultured and wild *Pseudo-nitzschia* (Bacillariophyceae) using species-specific LSU rRNA-targeted fluorescent probes and filter-based whole cell hybridization. *J. Phycol.* **34**, 371–382 (1998).
- Scholin, C. A., Miller, P. E., Buck, K. R. & Chavez, F. P. Detection and quantification of *Pseudo-nitzschia australis* in cultured and natural populations using LSU rRNA-targeted probes. *Limnol. Oceanogr.* **42**, 1265–1272 (1997).
- Scholin, C. A. *et al.* Application of DNA probes and a receptor binding assay for detection of *Pseudo-nitzschia* (Bacillariophyceae) species and domoic acid activity in cultured and natural samples. *J. Phycol.* (in the press).
- Van Dolah, F. M., Leighfield, T. A., Haynes, B. L., Hampson, D. R. & Ramsdell, J. S. A microplate receptor assay for the amnesic shellfish poisoning toxin, domoic acid, utilizing a cloned glutamate receptor. *Anal. Biochem.* **245**, 102–105 (1997).
- Parsons, T. R., Takahashi, M. & Hargrave, B. (eds) *Biological Oceanographic Processes*, 3rd edn, 39 (Pergamon, New York, 1984).
- Chavez, F. P. *et al.* Horizontal transport of and the distribution of nutrients in the coastal transition zone off Northern California: effects on primary production, phytoplankton biomass and species composition. *J. Geophys. Res.* **96**, 14833–14848 (1991).
- Bates, S. S., Garrison, D. L. & Horner, R. A. in *The Physiological Ecology of Harmful Algal Blooms* (eds

- Anderson, D. M., Hallegraeff, G. M. & Cembella, A. D.) 267–292 (NATO Advanced Study Institute Series, Springer, Heidelberg, 1998).
16. Bates, S. S. in *The Physiological Ecology of Harmful Algal Blooms* (eds Anderson, D. M., Hallegraeff, G. M. & Cembella, A. D.) 405–426 (NATO Advanced Study Institute Series, Springer, Heidelberg, 1998).
 17. Gulland, F. et al. *Unusual Marine Mammal Mortality Event—Domoic Acid Toxicity in California Sea Lions (*Zalophus californianus*) Stranded Along the Central California Coast, May–October 1998* (NOAA Technical Memo NMFS-OPR-8, National Marine Fisheries Service, US Department of Commerce, Silver Spring, Maryland, 1999).
 18. Sutherland, R. J., Hoising, J. M. & Whisaw, I. Q. Domoic acid, an environmental toxin, produces hippocampal damage and severe memory impairment. *Neurosci. Lett.* **120**, 221–223 (1990).
 19. Strain, S. M. & Tasker, R. A. R. Hippocampal damage produced by systemic injections of domoic acid in mice. *Neuroscience* **44**, 343–352 (1991).
 20. Dakshinamurti, K., Sharma, S. K., Sundaram, M. & Wanatabe, T. Hippocampal changes in developing postnatal mice following intrauterine exposure to domoic acid. *J. Neurosci.* **13**, 4486–4495 (1993).
 21. Antonelis, G. A., Fiscus, C. H., DeLong, R. H. Spring and summer prey of California sea lions, *Zalophus californianus*, at San Miguel Island, California, 1978–1979. *Fish. Bull.* **82**, 67–76 (1984).
 22. Ochoa, J. I., Sierra-Beltrán, A. P., Oláiz-Fernández, G. & Del Villar Ponce, L. M. Should mollusc toxicity in Mexico be considered a public health issue? *J. Shellfish Res.* **17**, 1671–1673 (1998).
 23. Sakamoto, C., Freiderich, G. E. & Codispoti, L. A. *MBARI Procedures for Automated Nutrient Analyses Using a Modified Alpkem Series 300 Rapid Flow Analyzer* (MBARI Technical Report 90-2, Monterey Bay Aquarium Research Institute, Moss Landing, California, 1990).
 24. Quilliam, M. A., Xie, M. & Hardstaff, W. R. A rapid extraction and clean-up procedure for the liquid chromatographic determination of domoic acid in unsalted seafood. *J. AOAC Int.* **78**, 543–554 (1995).
 25. Hatfield, C. L., Wekell, J. C., Gauglitz, E. J. Jr & Barnett, H. J. Salt clean-up procedure for the determination of domoic acid by HPLC. *Nat. Toxins* **2**, 206–211 (1994).
 26. McClain, C. R. et al. Science quality seaWiFS data for global biosphere research. *Sea Technol.* **39**, No. 9 (1998).
 27. Luna, L. G. *Manual of Histologic Staining Methods of the Armed Forces Institute of Pathology*, 3rd edn (McGraw-Hill, New York, 1968).

Acknowledgements

We thank the many volunteers who assisted The Marine Mammal Center (TMMC), members of the Monterey Bay Sanctuary Beach COMBERS programme, the crew of the RV *Pt. Lobos*, R. Kudela for providing the processed SeaWiFS image, and P. Walz for nutrient analyses. This work was supported in part by NOAA/NOS operational funds, NOAA Saltonstall-Kennedy grant programme, the many supporters of TMMC, and The David and Lucile Packard Foundation. All authors contributed equally to the work.

Correspondence and requests for materials should be addressed to C.S.
(e-mail: scholin@mbari.org).

Dynamic biogeography and conservation of endangered species

Rob Channell* & Mark V. Lomolino†

* Department of Biological Sciences, Fort Hays State University, Hays, Kansas 67601, USA

† Oklahoma Biological Survey, Oklahoma Natural Heritage Inventory, and Department of Zoology, University of Oklahoma, Norman, Oklahoma 73019, USA

As one moves from the core to the periphery of a species' geographical range, populations occupy less favourable habitats and exhibit lower and more variable densities^{1–4}. Populations along the periphery of the range tend to be more fragmented and, as a result, are less likely to receive immigrants from other populations. A population's probability of extinction is directly correlated with its variability and inversely correlated with density and immigration rate^{5–9}. This has led to the prediction that, when a species becomes endangered, its geographical range should contract inwards, with the core populations persisting until the final stages of decline^{2,10}. Convinced by these logical but untested deductions, conservation biologists and wildlife managers have been instructed to avoid the range periphery when planning conservation strategies or allocating resources for endangered species^{11–13}. We have analysed range contraction in 245 species from a broad range of taxonomic groups and geographical regions. Here we report that observed patterns of

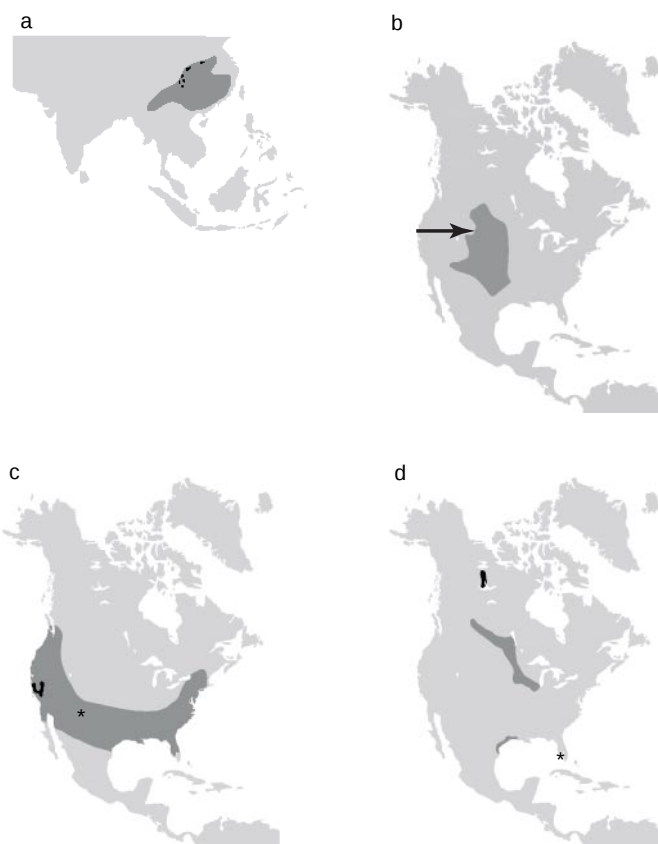


Figure 1 Patterns of range contraction in four endangered species. **a**, Giant panda, *Ailuropoda melanoleuca*; **b**, black-footed ferret, *Mustela nigripes*; **c**, California condor, *Gymnogyps californianus*; **d**, whooping crane, *Grus americana*. Historical range is in grey, extant range is in black or indicated by an arrow, and asterisks mark the locations of recent re-introduction sites for the California condor and the whooping crane.

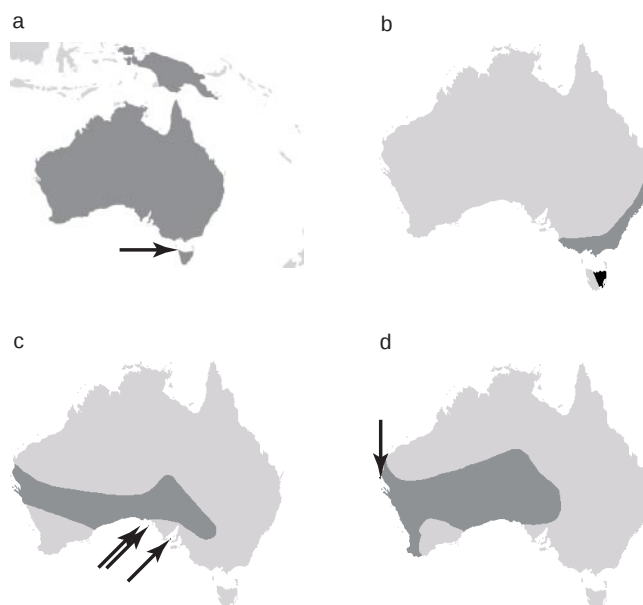


Figure 2 Patterns of range contraction in four species whose historical range included islands as well as much larger areas on the Australian mainland. **a**, Tasmanian tiger, *Thylacinus cynocephalus*; **b**, Tasmanian bettong, *Bettongia gaimardi*; **c**, greater stick-nest rat, *Leporillus conditor*; **d**, Shark Bay mouse, *Pseudomys fieldi*. Historical range in grey, and extant or final range is in black or indicated by an arrow.

Table 1 Number of species studied from different taxonomic groups and geographical regions

	North America	Australia	Eurasia	South America	Africa	Islands	Subtotal
Birds	12	6	19	2	3	45	87
Mammals	8	36	30	5	20	1	100
Reptiles	1		2	1		1	5
Amphibians	3		1				4
Fishes	1		1				2
Mollusks	1		1			20	22
Arthropods	2		1		1		4
Plants	4					17	21
Subtotal	32	42	55	8	24	84	245

See ref. 24.

range contraction do not support the above predictions and that most species examined persist in the periphery of their historical geographical ranges.

Table 1 shows the number of species studied and their geographical distribution. We found that 240 (98%) of the 245 species maintained populations in at least a portion of their peripheral range. Furthermore, 167 (68%) maintained a greater than expected portion of their range in the periphery, not the core ($P < 0.001$, binomial test). In fact, remnant populations of 91 species occurred exclusively in the periphery of their historical range, whereas populations of only five species persisted solely in the core of their historical range ($P < 0.001$, binomial test). We detected no significant difference in the patterns of range contraction between birds and mammals (63 (72%) of 87 birds and 70 (70%) of 100 mammals exhibiting greater persistence along the periphery). Most species, including some of the flagship species of conservation biology (Fig. 1), persist along the edge of their range.

Consistent with contemporary theory in ecology^{6,7,9}, persistence was greater for populations occupying larger patches of their historical range. On the mainland, 12 (75%) of 16 species persisted in larger patches of their historical range, whereas 15 (83%) of 18 insular species persisted in larger patches. However, if a species' historical range included both mainland and insular sites, population persistence was highest on the islands, despite their smaller size (23 [68%] of 34 species exhibited greater than expected persistence on islands; $P = 0.029$, binomial test; Fig. 2).

We found two additional patterns that seem contrary to the general tendency for greater persistence along the range periphery—Africa and the Hawaiian Islands. Africa was the only continent with an adequate sample size whose species failed to exhibit a significant peripheral bias in persistence (14 (58%) of 24 species persisted in the periphery; $P = 0.271$, binomial test). In contrast, 42 (78%) of 54 Eurasian species, 34 (81%) of 42 Australian species and 26 (81%) of 32 North American species persisted in their range peripheries ($P < 0.001$, 0.001, 0.001, respectively, binomial tests). In a similar fashion, whereas 11 (92%) of the 12 species we studied from New Zealand, and all of the 6 species from the Mariana Islands (including Guam) persisted more in the periphery than expected by chance, only 43% of the 54 Hawaiian species exhibited a peripheral bias.

These apparently exceptional results and the more general tendency for persistence along the periphery indicate that range contraction is strongly influenced by anthropogenic extinction forces (for example, habitat degradation, biocides and introduced species) which render historical density patterns irrelevant. Populations that persist the longest are those last affected by the contagion-like spread of extinction forces; that is, those along the edge of the range, on an isolated and undisturbed island, or at high elevations. African species failed to show any peripheral bias in range decline because, instead of moving across species' geographical ranges like a contagion, humans having a significant ecological effect became established in many places across the continent before the earliest record of historical extinctions. We actually predicted this result for Africa, based largely on Martin's^{14,15} explanation for the absence of a

post-Pleistocene collapse of the African megafauna: large mammals and birds shared a long evolutionary and ecological history with prehistoric humans. The 'exceptional' pattern for Hawaiian species is also entirely consistent with the above hypothesis concerning the contagion-like spread of extinction forces. Polynesians and, later, Europeans colonized most of the beach front and lowlands of these islands, and then spread, along with their commensals, upward. Persistent populations of Hawaiian species are either those that can cope with these anthropogenic disturbances, or those whose final populations remain in the least disturbed and most isolated sites; that is, in the montane areas. In short, the geography of recent extinctions is largely the geography of humanity. Thus, our ability to understand patterns in recent extinctions and to predict those of future ones depends to a very large degree on our ability to reconstruct and predict the spatial dynamics of humans and associated extinction forces.

These results have strong implications for conservation biology. Although they may have represented suboptimal habitats in historical times, areas along the range periphery and on remote islands and mountain ranges often provide valuable opportunities for conserving endangered species^{16,20}. We find it very encouraging, therefore, that a number of recent conservation programmes have broadened their options by including peripheral sites for re-introductions and areas to search for undiscovered populations of endangered species (asterisks in Fig. 1c, d). Although once viewed as the land of the living dead^{21,22}, sites along the range periphery may now hold great promise for conserving endangered species and biological diversity in general. □

Methods

We obtained range maps for 245 species from the literature or through personal correspondence with authorities (see Supplementary Information). We include only those species with maps available for both historical and extant ranges (or final site in the case of extinct species), and with extant ranges that were less than 25% of the species' historical distribution. We digitized the range maps into Idrisi, a geographical information system²³. For each species, we first located the centre, which was the point within the species' historical range that was most distant from all edges of the range. The distance from this point to the nearest edge was then calculated. We defined the region that was within half of this distance to an edge as periphery and the remaining portion of the range as central. We then calculated an index of centrality (C), which is a measure of the proportion of the extant or final range that fell within the central region of the historical range.

First, we calculated the area of the extant range expected to occur within the central region (C_{EE}) as follows:

$$C_{EE} = \left(\frac{C_H}{T_H} \right) T_E,$$

where T_E is the total area of the extant (or final) range; T_H is the total area of the historical range; and C_H is the area of the central region of the historical range. We then calculated C as follows. If $C_{EO} \leq C_{EE}$, where C_{EO} is the area of the extant range observed within the historical central region, then

$$C = \left(\frac{C_{EO}}{C_{EE}} \right) 0.5$$

If $C_{EO} > C_{EE}$, then

$$C = 0.5 + \left[0.5 \left(\frac{C_{EO} - C_{EE}}{T_E} \right) \right].$$

The index of centrality (C) ranged from 0, where the extant range fell completely outside the central portion of the historical range, to 1, where the extant range fell completely within the central portion of the historical range. We designated species with C values greater than 0.5 as 'central species', and those species with C values less than 0.5 as 'peripheral species'. We then used a binomial test to determine whether the ratio of central to peripheral species differed significantly from 1 : 1.

We used maps for species with multiple patches in their historical range to test whether persistence was higher for populations inhabiting larger patches. We first assigned patches to one of two size categories ('large' or 'small'), based on their area relative to the median patch size. If a species had an odd number of patches in its historical range, the median-sized patch was excluded from the analysis. For each species, we counted the number of large and small patches maintaining persistent populations (P_L and P_S , respectively). We counted the number of species (S_L) for which P_L was greater than P_S and the number of species (S_S) where P_S was greater than P_L . Species with ties ($P_L = P_S$) were excluded from analysis. We used a binomial test to determine whether the ratio of S_L to S_S differed significantly from 1 : 1. This analysis was done for 124 continental and 44 insular species²⁴.

To compare the relative persistence of mainland and island patches, we first calculated the total area of all of the historical patches (A_{TH}) and the area of the historical mainland patches (A_{MH}) for 44 species. We multiplied A_{MH}/A_{TH} by the total number of persisting patches (P_{TP}) to generate the expected number of patches persisting on the mainland. If the number of patches persisting on the mainland (P_{MP}) was greater than expected, we classified the species as a mainland species, otherwise it was classified as an island species. There were no ties ($P_{MP} = \text{expected number of patches}$). We tested whether the ratio of mainland species and island species differed significantly from 1 : 1 using a binomial test.

Received 23 April; accepted 21 October 1999.

1. Brown, J. H. On the relationship between abundance and distribution of species. *Am. Nat.* **124**, 255–279 (1984).
2. Lawton, J. H. In *Extinction Rates* (eds Lawton, J. H. & May, R. M.) 147–163 (Oxford Univ. Press, 1995).
3. Brown, J. H., Mehlman, D. W. & Stevens, G. C. Spatial variation in abundance. *Ecology* **76**, 2028–2043 (1995).
4. Gaston, K. J. Patterns in the geographical ranges of species. *Biol. Rev.* **65**, 105–129 (1990).
5. MacArthur, R. H. & Wilson, E. O. The theory of island biogeography. *Monogr. Popul. Biol.* **1**, 1–203 (1967).
6. Pimm, S. L., Jones, H. L. & Diamond, J. On the risk of extinction. *Am. Nat.* **132**, 757–785 (1988).
7. Tracy, C. R. & George, T. L. On the determinants of extinction. *Am. Nat.* **139**, 102–122 (1992).
8. Brown, J. H. & Kodric-Brown, A. Turnover rates in insular biogeography: effects of immigration on extinction. *Ecology* **58**, 445–449 (1977).
9. Goel, N. S. & Richter-Dyn, N. *Stochastic Models in Biology* (Academic, New York, 1974).
10. Brown, J. H. *Macroecology* (Univ. Chicago Press, 1995).
11. Wolf, C. M., Griffith, B., Reed, C. & Temple, S. A. Avian and mammalian translocations: update and reanalysis of 1987 survey data. *Conserv. Biol.* **10**, 1142–1153 (1996).
12. Griffith, B., Scott, J. M., Carpenter, J. W. & Reed, C. Translocation as a species conservation tool: status and strategy. *Science* **245**, 477–480 (1989).
13. Pearl, M. In *Conservation Biology: The Theory and Practice of Nature Conservation Preservation and Management* (eds Fielder, P. L. & Jain, S. K.) 297–320 (Chapman & Hall, New York, 1992).
14. Martin, P. S. In *Quaternary Extinctions* (eds Martin, P. S. & Klein, R. G.) 354–403 (Univ. Arizona Press, Tucson 1984).
15. Diamond, J. M. In *Quaternary Extinctions* (eds Martin, P. S. & Klein, R. G.) 824–862 (Univ. Arizona Press, Tucson, 1984).
16. Burbidge, A. A. & McKenzie, N. L. Patterns in the modern decline of western Australia's vertebrate fauna: causes and conservation implications. *Biol. Conserv.* **50**, 143–198 (1989).
17. Short, J., Bradshaw, S. D., Giles, J., Prince, R. I. T. & Wilson, G. R. Reintroduction of macropods (Marsupialia: Macropodoidea) in Australia—A review. *Biol. Conserv.* **62**, 189–204 (1992).
18. Bibby, C. J. Recent past and future extinctions in birds. *Phil. Trans. R. Soc. Lond. B* **344**, 35–40 (1994).
19. Franklin, J. & Steadman, D. W. The potential for conservation of Polynesian birds through habitat mapping and species translocation. *Conserv. Biol.* **5**, 506–521 (1991).
20. Towns, D. R. & Daugherty, C. H. Patterns of range contractions and extinctions in the New Zealand herpetofauna following human colonisation. *N.Z. J. Zool.* **21**, 325–339 (1994).
21. Stevens, G. In *Systematics, Ecology, and the Biodiversity Crisis* (ed. Eldredge, N.) 40–58 (Columbia Univ. Press, New York, 1992).
22. Wolf, C. M., Griffith, B., Reed, C. & Temple, S. A. Avian and mammalian translocations: update and reanalysis of 1987 survey data. *Conserv. Biol.* **10**, 1142–1153 (1996).
23. Eastman, J. R. *Idrisi for Windows, Version 1.0* (Clark Labs for Cartographic Technology and Geographic Analysis, Worcester, MA, 1995).
24. Channell, R. A geography of extinction: patterns in the contraction of geographic ranges. Thesis, Univ. Oklahoma (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial offices of Nature.

Acknowledgements

We thank A. Baynes, J. H. Brown, N. Czaplewski, B. Danielson, T. Franklin, M. Kaspari, B. Maurer, K. Pandora, D. Perault, K. Perez, S. Pimm, G. A. Smith and C. Vaughn for advice and comments on this paper, and J. M. Scott, D. Steadman and L. Carbyn for information on the distribution of several species. R.C. was supported by the Department of Zoology, University of Oklahoma, while conducting this research, and M.V.L. was supported by grants from the US National Science Foundation.

Correspondence and requests for materials should be addressed to R.C. (e-mail: Rchannel@FHSU.edu).

Reduced vas deferens contraction and male infertility in mice lacking P2X₁ receptors

K. Mulryan*, D. P. Gitterman*, C. J. Lewis*, C. Vial*, B. J. Leckie†, A. L. Cobb‡, J. E. Brown‡, E. C. Conley§, G. Buell||, C. A. Pritchard¶ & R. J. Evans*

* Department of Cell Physiology & Pharmacology, Medical Sciences Building,

† Department of Medicine, ‡ Transgenic Unit, Biomedical Services,

§ Department of Pathology and Centre for Mechanisms of Human Toxicity &

¶ Department of Biochemistry, University of Leicester, Leicester LE1 9HN, UK

|| Glaxo-Wellcome Geneva Biomedical Research Institute, 14 chemin des Aulx, Plan-les-Ouates, 1228 Geneva, Switzerland

Present address: SeroPharmaceuticals, 14 chemin des Aulx, Plan-les-Ouates, 1228 Geneva, Switzerland

P2X₁ receptors for ATP are ligand-gated cation channels, present on many excitable cells including vas deferens smooth muscle cells^{1–5}. A substantial component of the contractile response of the vas deferens to sympathetic nerve stimulation, which propels sperm into the ejaculate, is mediated through P2X receptors¹. Here we show that male fertility is reduced by ~90% in mice with a targeted deletion of the P2X₁ receptor gene. Male mice copulate normally—reduced fertility results from a reduction of sperm in the ejaculate and not from sperm dysfunction. Female mice and heterozygote mice are unaffected. In P2X₁-receptor-deficient mice, contraction of the vas deferens to sympathetic nerve stimulation is reduced by up to 60% and responses to P2X

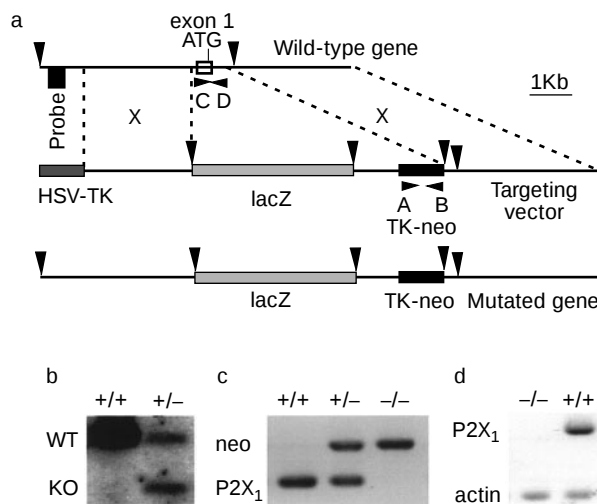


Figure 1 Generation of P2X₁-receptor-deficient mice. **a**, Genomic maps of the wild-type gene, targeting vector and mutated gene. BamHI sites (indicated by arrows) and the probe used for detection of the homologous recombination events by Southern analysis are shown. Polymerase chain reaction (PCR) primers used for genotyping of mouse-tail DNA are indicated (A–D). **b**, Southern blot analysis of tail genomic DNA from +/+ and -/- animals. Genomic DNA was digested with BamHI and hybridized with the probe indicated in **a** which detects a 4.8-kb band in +/+ DNA and a 3.7-kb band in -/- DNA. WT, wild-type; KO, knock-out. **c**, PCR genotyping of mouse-tail DNA. Primers A, B, C and D were used in one PCR reaction to genotype mouse-tail genomic DNA. Primers A and B amplify a 519-bp product from the *neo*^R gene, whereas primers C and D amplify a 317-bp product from the deleted region of the P2X₁ receptor gene. **c**, RT-PCR analysis. A PCR product of 442 bp from the P2X₁-receptor gene was amplified from bladder complementary DNA from a +/+ animal but not from bladder cDNA of a -/- animal. As a control, amplification of 199-bp product from the actin gene was detected in both samples.

receptor agonists are abolished. These results show that P2X₁ receptors are essential for normal male reproductive function and suggest that the development of selective P2X₁ receptor antagonists may provide an effective non-hormonal male contraceptive pill. Also, agents that potentiate the actions of ATP at P2X₁ receptors may be useful in the treatment of male infertility.

ATP is co-released with noradrenaline from sympathetic nerves and acts through P2X receptors on smooth muscle to mediate membrane depolarization and contraction¹. Seven P2X receptors have been identified at the molecular level¹. P2X receptors were originally isolated from the rat vas deferens and can account for the rapidly inactivating α,β -methylene ATP (α,β -meATP)-sensitive native vas deferens P2X receptor phenotype⁶. In smooth muscle the P2X₁ receptor is the dominant form expressed⁵. Owing to the lack of potent and subtype-selective P2X-receptor antagonists the physiological role of P2X₁ receptors has been difficult to determine. To overcome these problems we have generated a P2X₁-receptor-deficient mouse.

The P2X₁-receptor-targeting vector was designed to remove the first 45 amino acids of the P2X₁ receptor (exon 1) and results in the functional inactivation of the P2X₁ receptor gene (Fig. 1). Male and female mice heterozygous for the P2X₁ receptor deficiency (+/-) are phenotypically normal and when inter-crossed produced progeny with a mendelian genotype distribution of 1+/+ : 2.1+/- : 1-/- (sex ratio 1.05 male:1 female, $n = 320$) indicating that there is no selective fertilization or mortality *in utero*. Confirmation of the deficiency of the P2X₁ receptor in homozygous mutant mice was obtained at the messenger RNA and protein levels. Polymerase chain reaction after reverse transcription of RNA (RT-PCR), using primers specific for a region 3' to that deleted by the targeting vector, amplified a 442-base pair (bp) product from +/+ mice but not from -/- mice, indicating that mRNA for a truncated form of the P2X₁ receptor is not produced by -/- mice (Fig. 1d). A P2X₁ receptor antibody raised against the carboxy terminus of the protein produced high levels of immunoreactivity in the smooth muscle layer of +/+ vas deferens as reported previously⁷. However, no P2X₁ receptor immunoreactivity was detected in the vas deferens of -/- mice (Fig. 2a and b).

Interbreeding of mutant -/- mice did not result in pregnancy ($n = 8$) even though coitus had taken place. Mutant -/- female mice produced litters when mated with +/+ ($n = 5$) or +/- males ($n = 3$). When -/- males were mated with +/+ females only 13.7%

of matings resulted in pregnancy ($n = 51$ matings of 22 -/- males) compared to a 100% pregnancy rate for +/- males ($n = 12$). These results indicate that the reduced fertility was specific to the male P2X₁-receptor-deficient mice. The mean litter size was also reduced for -/- males by 30% (mean litter size -/- 6.6 ± 1.0 , $n = 6$; +/- 9.5 ± 0.46 , $n = 12$; $P < 0.05$). Taking together the reduction in pregnancy rate and litter size, the fecundity of -/- males is reduced by 90.4%.

A vaginal coagulum/plug is formed following copulation. This results from the action of enzymes in the seminal plasma secreted from the male sex accessory glands. Coagulum weight was essentially the same following mating with +/+ or +/- mice (47.2 ± 2.7 mg, $n = 5$) or -/- males (57.3 ± 3.7 mg, $n = 7$) indicating that similar ejaculate volumes were produced. Given that the fluids secreted from accessory sexual glands account for ~80% of total semen volume these findings suggest that the contraction and function of these glands is unaffected in the P2X₁-receptor-deficient mouse. The reduction in male fertility could result from problems with spermatogenesis or sperm quality and/or the reduction of sperm numbers in the ejaculate. P2X₁ receptor immunoreactivity was detected in blood vessels in the +/+ testis; however, no immunoreactivity was detected in the seminiferous tubules (Fig. 2c), indicating that P2X₁ receptors are not involved in spermatogenesis. In addition, haematoxylin- and eosin-stained sections of testis from -/- mice appeared normal. Vas deferens from +/+, +/- and -/- mice appeared equally full of sperm. Motile sperm were recovered from +/+, +/- and -/- epididymis and vas deferens in similar numbers (epididymal count $(3.6 \pm 0.8) \times 10^6 \text{ ml}^{-1}$ for +/+ and +/-, $(5.1 \pm 1.6) \times 10^6 \text{ ml}^{-1}$ for -/-, $n = 5-6$) and were equally effective at fertilizing ova *in vitro* (Fig. 2b and d, % success at 4 days $70.0\% \pm 2.6\%$ for +/+ and $65.5\% \pm 5.9\%$ for -/-, $n = 3$). Thus infertility does not result from

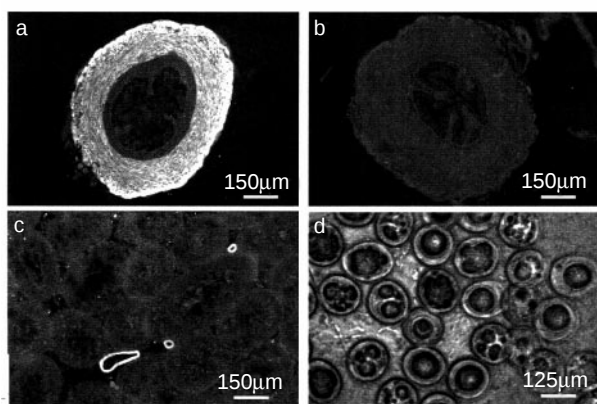


Figure 2 Immunohistochemical detection of P2X₁ receptors and *in vitro* fertilization studies. **a, b**, Confocal images of transverse sections of vas deferens show immunohistochemical detection of the P2X₁ receptor protein in the smooth-muscle layers of the wild-type vas deferens (**a**); no immunoreactivity was present in the -/- mutant animal (**b**). **c**, Confocal image of a transverse section of testis shows P2X₁-receptor immunoreactivity associated with blood vessels, but no reactivity with the seminiferous tubules. **d**, Fertilization of ova *in vitro* by sperm extracted from -/- epididymis, photomicrograph taken at 4 days.

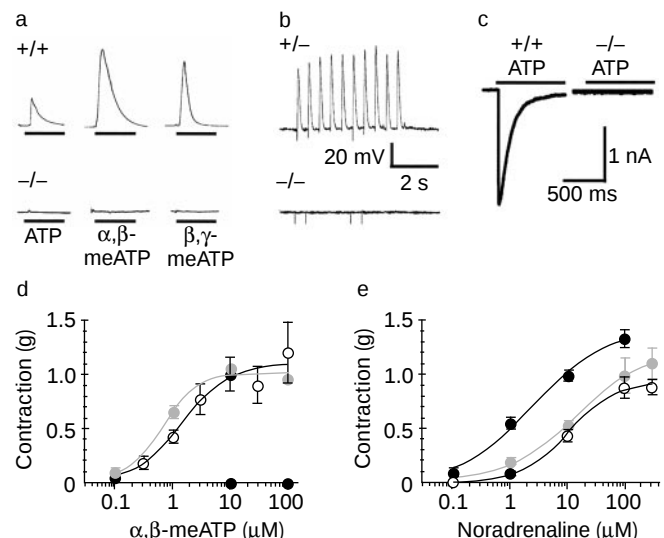


Figure 3 Response to purinergic agonists, nerve stimulation and noradrenaline of wild-type and P2X₁-receptor-deficient mouse vas deferens. **a**, ATP (100 μM), α,β -meATP and 1- β,γ -meATP (both 10 μM) evoked transient contractions of +/+ vas deferens but had no effect on the tone of -/- vas deferens. **b**, Intracellular membrane potential recordings from vas deferens smooth-muscle cells. Sympathetic nerve stimulation (10 pulses at 10 Hz, 0.5 ms pulse width) evoked excitatory junction potentials in +/- but not -/- vas deferens (resting membrane potential -87.7 ± 1 mV and 87.6 ± 1.7 mV for +/- and -/- mice respectively, $n = 16$ for each). **c**, Whole-cell patch-clamp recordings from acutely dissociated vas deferens smooth muscle cells. ATP evoked a rapidly inactivating inward current in +/+ but had no effect on -/- acutely dissociated vas deferens smooth muscle cells (holding potential -60 mV). (**a, c**, Agonist applications indicated by bar, 60s for **a**). **d, e**, Concentration-effect relationship for α,β -meATP (**d**) and noradrenaline (**e**) for +/+ (open circles), +/- (shaded circles) and -/- (solid circles) vas deferens segments.

a sperm dysfunction. Similarly reduced fertility is unlikely to have been caused by blockage of the vas deferens as this results in degeneration of sperm in the epididymis⁸. Analysis of lavage from the uterus, taken at the junction with the coagulum, showed that sperm were abundant in mice mated with $+/+$ ($n = 3$) or $+/-$ mice ($n = 2$), but no sperm could be detected in the lavage from matings with $-/-$ males ($n = 7$). These results show that the reduction in $-/-$ male fertility results from a reduced sperm count.

Firing of sympathetic nerves associated with ejaculation results in contraction of the vas deferens and emission of sperm into the semen. A substantial component of this sympathetic-nerve-evoked contraction is mediated through the activation of P2X receptors⁹. P2X₁-receptor-deficient mice may therefore be compromised in vas deferens function and this could explain the absence of sperm in the ejaculate. In the $+/+$ vas deferens the purinergic agonists ATP, α,β -meATP and 1- β,γ -meATP evoked transient contractions through the activation of P2X receptors, which declined to baseline during the continued presence of the agonist as described previously^{10,11} (Fig. 3a). The pEC_{50} values of α,β -meATP for evoking contraction of the vas deferens for $+/+$ and $+/-$ mice are 5.83 ± 0.05 and 5.72 ± 0.04 , respectively (Fig. 3d). P2X receptors are thought to underlie the excitatory junction potentials (EJPs) recorded from smooth muscle¹². Spontaneous EJPs and EJPs evoked by sympathetic nerve stimulation were recorded from all impalements of the $+/-$ vas deferens ($n = 16$ impalements from four vasa deferentia) (Fig. 3b). In patch clamp studies on acutely dissociated $+/+$ vas deferens smooth muscle cells α,β -meATP and ATP evoked transient inward currents (amplitude 2861 ± 410 pA, $n = 8$ and 2096 ± 175 pA, $n = 3$, respectively) with latency <10 ms and with 10%–90% rise times of 22.7 ± 7.7 ms ($n = 4$), demonstrating the direct activation of a ligand-gated P2X-receptor channel^{3,11} (Fig. 3c). In contrast, the purinergic agonists α,β -meATP and ATP failed to evoke contractions (Fig. 3a, d) or inward currents (Fig. 3c) and spontaneous or evoked EJPs were not detected (Fig. 3b) ($n = 16$ impalements from four vasa deferentia) from the vas deferens of the $-/-$ P2X₁-receptor-deficient mouse. It has been suggested, on the

basis of contraction studies, that there may be multiple P2X receptor subtypes in the vas deferens (see references in ref. 13). In the present study vas deferens P2X receptors showed the hallmark properties of homomeric P2X₁ receptors (transient α,β -meATP and 1- β,γ -meATP sensitive responses). The lack of functional P2X receptors in the $-/-$ vas deferens shows that if subtypes do exist the expression of P2X₁ receptors is essential for the production of functional receptors.

The lack of contraction of the $-/-$ vas deferens to purinergic agonists does not result from an impairment of the contractile function, as responses evoked by 100 mM potassium chloride were the same for $+/+$ and $-/-$ P2X₁-receptor-deficient mice (1.36 ± 0.12 g and 1.17 ± 0.08 g, respectively; $n = 14$ –17). In addition, $+/+$, $+/-$ and $-/-$ vas deferens contracted in response to noradrenaline (Fig. 3e). The vas deferens of the $-/-$ mice were more sensitive to noradrenaline than the $+/+$ (pEC_{50} values of 5.65 ± 0.15 and 4.79 ± 0.09 , respectively, $P = 0.005$, $n = 4$ –6) and the $+/-$ mice had an intermediate sensitivity with a pEC_{50} of 5.1 ± 0.27 ($n = 5$). The increase in sensitivity to noradrenaline in $-/-$ animals suggests a compensatory change in $-/-$ animals which could be accounted for by an increase in α_1 -adrenoceptor numbers¹⁴.

The lack of P2X receptor-mediated contractions in $-/-$ vas deferens suggests that contractions in response to sympathetic nerve stimulation may also be affected. In rats the duration of copulation from the first pelvic thrust to dismount is 1–3 s (ref. 15). To determine the response of the vas deferens to nerve stimulation the vas deferens was stimulated with a train of 60 pulses at 20 Hz (train duration 3 s) (Fig. 4). Electrically evoked responses were mediated through the activation of sympathetic nerves, as they were abolished by tetrodotoxin ($0.3 \mu\text{M}$) or the adrenergic neuron blocker guanethidine ($3 \mu\text{M}$) ($n = 4$). The peak amplitude of contraction of $+/+$ mouse vasa deferentia was 1.16 ± 0.08 g ($n = 21$). Desensitization of P2X receptors with α,β -meATP revealed that the P2X-receptor-mediated contraction accounted for $79 \pm 5.5\%$ of the response ($n = 7$). The residual response was abolished by the α_1 -adrenoceptor antagonist prazosin (Fig. 4a and c). Similar responses were recorded from $+/-$ males (Fig. 4c). In contrast, the peak amplitude of contractions of $-/-$ vas deferens was only $\sim 60\%$ (0.7 ± 0.04 g, $n = 36$) of the $+/+$ response ($n = 16$), was reduced by $<10\%$ by α,β -meATP treatment and was abolished by prazosin ($n = 4$) (Fig. 4b and d). This increase in magnitude of the α_1 -adrenoceptor-mediated component in $-/-$ mice, like the response to exogenously applied noradrenaline, can be accounted for by an increase in α_1 -adrenoceptor number¹⁴. Long trains of high-frequency stimulation are known to favour the noradrenergic-mediated component of smooth-muscle contraction; however, they probably do not reflect the *in vivo* nerve traffic.

The physiological sympathetic nerve firing pattern resulting in ejaculation remains to be determined. It is likely that the burst of sympathetic nerve activity to the vas deferens associated with ejaculation is considerably shorter than the duration of copulation. We therefore determined the effects of 0.5-s trains of stimuli at different frequencies. The magnitude of the contractile response was frequency dependent (Fig. 4d) and in response to 10-Hz stimulation the response of $-/-$ vas deferens was only 40% of the $+/+$ response. A similar decrease in ejaculate sperm count results in infertility¹⁶. Thus, this reduction in the neurogenic vas deferens contraction can account for the decrease in male fertility rate and suggests that in the majority of cases, *in vivo* contraction of the $-/-$ vas deferens is below the threshold required to eject the sperm into the semen.

In addition to the vas deferens, P2X₁ receptors are also present on a variety of other smooth-muscle preparations, including the urinary bladder and arteries, as well as parts of the nervous system. There was no obvious effect on the behaviour of P2X₁-receptor $-/-$ mice, and heart rate (572.2 ± 20.3 beats per minute for $-/-$ and 556 ± 19 beats per minute for $+/+$, $n = 8$ for each), and bladder function all appeared normal. There was, however, a small

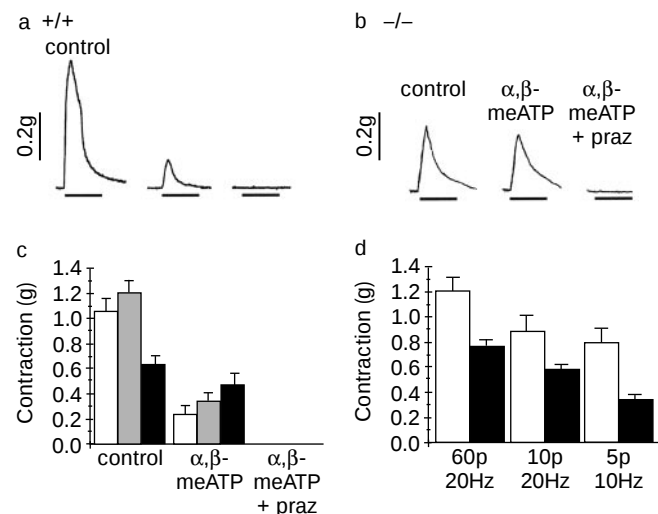


Figure 4 Sympathetic nerve-mediated contraction of $+/+$ and $-/-$ P2X₁-receptor-deficient vas deferens. **a**, **b**, Nerve stimulation, 60 pulses at 20 Hz (period indicated by bar) evoked contractions of $+/+$ (**a**) and $-/-$ (**b**) vas deferens. Following desensitization of P2X-receptor-mediated responses with α,β -meATP ($10 \mu\text{M}$), the residual contraction was abolished by the α_1 -adrenoceptor antagonist prazosin ($0.1 \mu\text{M}$). **c**, Summary of the effects of P2X-receptor-desensitization with α,β -meATP and the α_1 -adrenoceptor antagonist prazosin on neurogenic contractions of the mouse vas deferens; $+/+$ (white columns), $+/-$ (grey columns) and $-/-$ (black columns). **d**, Effects of stimulus frequency and duration on neurogenic contractions of the mouse vas deferens. Histogram shows peak amplitude of contraction to stimulation with 60 pulses at 20 Hz, 10 pulses at 20 Hz and 5 pulses at 10 Hz for $+/+$ (open columns) and $-/-$ (black columns) mice.

increase in systolic blood pressure at rest (115.8 ± 2 mm Hg for $-/-$ and 108 ± 1.8 mm Hg for $+/+$, as measured by the tail cuff method; $P = 0.016$, $n = 8$ for each).

We have shown that P2X₁ receptors are involved in the contraction of the vas deferens and that P2X₁ receptor deficiency results in a 90% decrease in male fertility through a reduction in sperm in the ejaculate associated with a decrease in neurogenic vas deferens contraction. In mice it appears that the residual α_1 -adrenoceptor-mediated neurogenic vas deferens contraction is insufficient for normal ejaculatory function. Selective α_1 -adrenoceptor antagonists do not cause azoospermia and infertility in man¹⁷ and contractile studies have indicated the presence of a substantial non-adrenergic component of contraction of prostatic portions of the human vas deferens¹⁸. This suggests that P2X₁-receptor antagonists may provide a target for the development of a non-hormonal male contraceptive pill. In addition, agents that potentiate the actions of ATP at P2X₁ receptors may be useful in the treatment of male infertility¹⁹. □

Methods

Generation of P2X₁ receptor-deficient mice

The targeting vector includes 5.8 kilobases (kb) of P2X₁-receptor genomic DNA, the *lacZ* gene, the *neo^R* gene driven by the TK promoter and the HSV-tk gene. Homologous recombination of this vector with the wild-type gene results in deletion of 350 bp of DNA which includes exon 1 and the initiating ATG. The deleted 350 bp are replaced by the *lacZ* gene and the *neo^R* gene of the targeting vector. The targeting vector was electroporated into E14.1a embryonic stem (ES) cells derived from the 129/Ola mouse strain, and colonies were selected with G418 and gancyclovir. Positive colonies were identified by the presence of the 3.7 kb band in *Bam*HI-digested genomic DNA. Four ES colonies containing the targeting event were microinjected into F1 (CBA $\times$ C57BL/6) blastocysts and chimaeras were derived. Germline transmission of the targeted allele was obtained from all four ES colonies upon mating to MF-1 animals. The mice analysed here have 129/Ola-MF-1 genetic backgrounds.

Immunohistochemistry

Detection of the distribution of the P2X₁ receptor using an antibody raised against the C terminus of the receptor was as described⁷. *In vitro* fertilization was as described²⁰.

Physiological studies

Mice were produced by crossing $+/-$ mice. Littermates were genotyped by PCR, animals were sexually mature (4–6 months old), weight was 37.5 ± 0.8 , 39.6 ± 1.3 and 37.3 ± 1.0 g for wild type, $+/-$ and $-/-$, respectively ($n = 15$ – 19). Mouse vasa deferentia were mounted in Ringers solution at 36°C in 15-ml organ baths under an initial load of 1 g, and tension was monitored isometrically. Agonists were applied to the bath at 30-min intervals and removed by washing; this solution was also used for intracellular recordings from the vas deferens using standard methods. Trains of electrical stimuli were delivered through silver chloride electrodes, 20–40 V 0.5 ms pulse width. Vas deferens smooth-muscle cells were enzymatically dissociated and responses to purinergic agonists applied rapidly by a U-tube perfusion system were determined in voltage-clamp recordings as described²¹.

Received 22 July; accepted 25 October 1999.

1. Burnstock, G. The past, present and future of purine nucleotides as signalling molecules. *Neuropharmacology* **36**, 1127–1139 (1997).
2. Surprenant, A., Buell, G. & North, R. A. P2X receptors bring new structure to ligand-gated ion channels. *Trends Neurosci.* **18**, 224–229 (1995).
3. Evans, R. J. & Surprenant, A. P2X receptors in autonomic and sensory neurons. *Semin. Neurosci.* **8**, 217–223 (1996).
4. Brading, A. F. & Williams, J. H. Contractile responses of smooth muscle strips from rat and guinea-pig urinary bladder to transmural stimulation: effects of atropine and α, β -methyleneATP. *Br. J. Pharmacol.* **99**, 493–498 (1990).
5. Collo, G. et al. Cloning of P2X₂ and P2X₃ receptors and the distribution and properties of an extended family of ATP-gated ion channels. *J. Neurosci.* **16**, 2495–2507 (1996).
6. Valera, S. et al. A new class of ligand-gated ion channel defined by P2X receptor for extracellular ATP. *Nature* **371**, 516–519 (1994).
7. Vulchanova, L. et al. Differential distribution of two ATP-gated ion channels (P2X receptors) determined by immunohistochemistry. *Proc. Natl Acad. Sci. USA* **93**, 8063–8067 (1996).
8. Costa, S. L., Boekelheide, K., Vanderhyden, B. C., Seth, R. & McBurney, M. W. Male infertility caused by epididymal dysfunction in transgenic mice expressing a dominant negative mutation of retinoic acid receptors alpha 1. *Biol. Reprod.* **56**, 985–990 (1997).
9. Allcorn, R. J., Cunnean, T. C. & Kirkpatrick, K. Actions of α, β -methylene ATP and 6-hydroxydopamine on sympathetic neurotransmission in the vas deferens of the guinea-pig, rat and mouse: support for co-transmission. *Br. J. Pharmacol.* **89**, 647–659 (1986).
10. von Kugelgen, I., Bultmann, R. & Starke, K. Interaction of adenosine nucleotides, UTP and suramin in mouse vas deferens: suramin-sensitive and suramin-insensitive components in the contractile effect of ATP. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **342**, 198–205 (1990).
11. Khakh, B. S., Surprenant, A. & Humphrey, P. P. A study on P2X purinoceptors mediating the electrophysiological and contractile effects of purine nucleotides in rat vas deferens. *Br. J. Pharmacol.* **115**, 177–185 (1995).

12. Burnstock, G. & Holman, M. E. The transmission of excitation from autonomic nerve to smooth muscle. *J. Physiol.* **155**, 115–133 (1961).
13. Ralevic, V. & Burnstock, G. Receptors for purines and pyrimidines. *Pharmacol. Rev.* **50**, 413–492 (1998).
14. Kenakin, T. in *Pharmacologic Analysis of Drug-Receptor Interaction* Vol. 3, 63–67 (Lippincott-Raven, Philadelphia, 1997).
15. Yang, L. & Clemens, L. G. Influence of male-related stimuli on female postejaculatory refractory period in rats. *Physiol. Behav.* **63**, 675–682 (1998).
16. Anderson, R. A., Oswald, C., Willis, B. R. & Zaneveld, L. J. D. Relationship between semen characteristics and fertility in electroejaculated mice. *J. Reprod. Fert.* **68**, 1–7 (1983).
17. Hoffman, B. B. & Lefkowitz, R. J. in *The Pharmacological Basis of Therapeutics* (eds Hardman, J. G., Limbird, L. E., Molinoff, P. B., Ruddon, R. W. & Gilman, A. G.) Vol. 9, 199–248 (McGraw-Hill, New York, 1996).
18. Hedlund, H., Andersson, K. & Larsson, B. Effect of drugs interacting with adrenoceptors and muscarinic receptors in the epididymal and prostatic parts of the human isolated vas deferens. *J. Auton. Pharmacol.* **5**, 261–270 (1985).
19. Jacobson, K. A. et al. A pyridoxine cyclic phosphate and its 6-azoaryl derivative selectively potentiate and antagonize activation of P2X₁ receptors. *J. Med. Chem.* **41**, 2201–2206 (1998).
20. Hogan, B., Constantini, F. & Lacy, E. *Manipulating the Mouse Embryo: A Laboratory Manual* (Cold Spring Harbour, New York, 1986).
21. Lewis, C., Surprenant, A. & Evans, R. J. 2',3'-O-(2,4,6-trinitrophenyl) adenosine 5'-triphosphate (TNP-ATP)—a nanomolar affinity antagonist at rat mesenteric artery P2X receptor ion channels. *Br. J. Pharmacol.* **1463**–**1466** (1998).

Acknowledgements

We thank L. Vulchanova and R. Elde for the P2X₁ receptor antibody, B. Grubb and S. Giblett for help with the immunohistochemistry and C. d'Lacey for help with the confocal images. We also thank J. Luckett and S. Monkley for help at the bench and S. R. Nahorski for comments on the manuscript. This work was supported by the Medical Research Council and the Wellcome Trust.

Correspondence and requests for materials should be addressed to R.J.E. (e-mail: RJE6@le.ac.uk).

Turning of nerve growth cones induced by localized increases in intracellular calcium ions

James Q. Zheng

Department of Neuroscience and Cell Biology, University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, New Jersey 08854, USA

Guidance of developing axons involves turning of the motile tip, the growth cone, in response to a variety of extracellular cues^{1,2}. Little is known about the intracellular mechanism by which the directional signal is transduced. Ca^{2+} is a key second messenger in growth cone extension^{3,4} and has been implicated in growth-cone turning^{5,6}. Here I report that a direct, spatially restricted elevation of intracellular Ca^{2+} concentration ($[\text{Ca}^{2+}]_i$) on one side of the growth cone by focal laser-induced photolysis (FLIP) of caged Ca^{2+} consistently induced turning of the growth cone to the side with elevated $[\text{Ca}^{2+}]_i$ (attraction). Furthermore, when the resting $[\text{Ca}^{2+}]_i$ at the growth cone was decreased by the removal of extracellular Ca^{2+} , the same focal elevation of $[\text{Ca}^{2+}]_i$ by FLIP induced repulsion. These results provide direct evidence that a localized Ca^{2+} signal in the growth cone can provide the intracellular directional cue for extension and is sufficient to initiate both attraction and repulsion. By integrating local and global Ca^{2+} signals, a growth cone could thus generate different turning responses under different environmental conditions during guidance.

To examine the response of *Xenopus* growth cones to spatially restricted cytosolic Ca^{2+} signals, FLIP of caged Ca^{2+} was used to elevate $[\text{Ca}^{2+}]_i$ directly in a spot $\sim 2\text{ }\mu\text{m}$ in diameter. The efficacy and spatial restriction of FLIP was first examined *in vitro* using

4',5-dimethoxy-2-nitrobenzyl-caged fluorescein-dextran. Simultaneous FLIP and fluorescence imaging showed that a single laser pulse (duration 4 ns, energy $\sim 10 \mu\text{J}$) liberated a significant amount of fluorescence from the caged fluorescein-dextran, indicating effective photoactivation (Fig. 1a). The spatial restriction of FLIP was clearly demonstrated when glycerol was used to slow down the diffusion of dextran molecules⁷ (Fig. 1a, upper panel); but without glycerol a much larger radius of fluorescence was observed (Fig. 1a, lower panel). This difference was evident even in the first frame of the sequence when analysed quantitatively (Fig. 1b, c). Next, direct focal elevation of $[\text{Ca}^{2+}]_i$ by FLIP was tested using the caged Ca^{2+} compound *o*-nitrophenyl EGTA (NP-EGTA)⁸. Simultaneous FLIP and Ca^{2+} imaging using fluo-3 showed that a single laser pulse induced a localized elevation of $[\text{Ca}^{2+}]_i$ in an NP-EGTA-loaded muscle cell, as shown by the spatially restricted increase in fluo-3 fluorescence (Fig. 1e); no change in $[\text{Ca}^{2+}]_i$ in the control muscle cell was induced by the laser (Fig. 1d). The localized elevation of $[\text{Ca}^{2+}]_i$ by FLIP was transient and could be detected only in the first frame of the image sequence. Similarly, in growth cones, FLIP of NP-EGTA also induced a spatially restricted elevation of $[\text{Ca}^{2+}]_i$ that was observed mostly in the first frame of the image sequence (Fig. 1f). Rapid diffusion as well as cytosolic buffering of the local Ca^{2+} signals^{9,10} probably contributes to this transient nature of the focally elevated $[\text{Ca}^{2+}]_i$. It should be noted that the imaging results do not accurately represent the focal $[\text{Ca}^{2+}]_i$ immediately after FLIP because rapid diffusion alone would probably result in the detection of smaller and more widespread local Ca^{2+} signals. Nevertheless, these results directly demonstrate the spatial restriction of FLIP for direct $[\text{Ca}^{2+}]_i$ elevation, which is of critical importance for the experiments below.

To determine whether a local change in $[\text{Ca}^{2+}]_i$ at the growth cone is sufficient to signal the direction of extension, $[\text{Ca}^{2+}]_i$ on one side

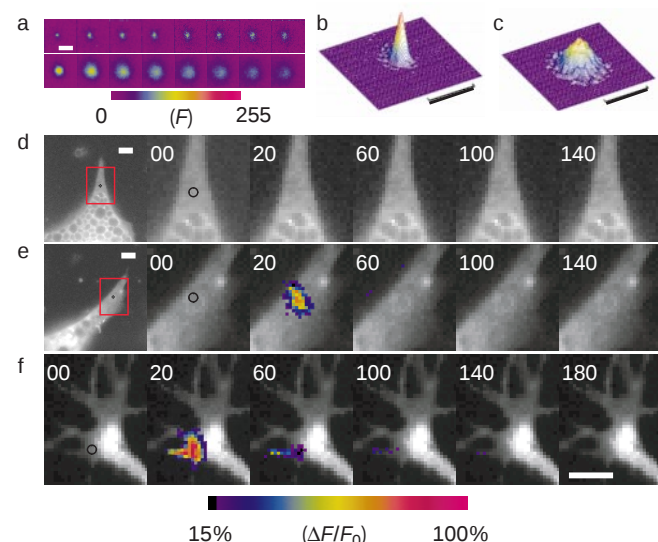


Figure 1 Spatially restricted FLIP of caged molecules. **a–c**, Spatiotemporal properties of fluorescence signals produced by FLIP of caged fluorescein-dextran *in vitro*. Two image sequences (**a**) demonstrate different diffusion rates of uncaged fluorescein-dextran in solutions with (upper) and without (lower) glycerol. The interval between each frame is 100 ms and the intensity of the fluorescence (F) is pseudo-colour-coded by using the look-up table illustrated in the colour bar shown. Three-dimensional intensity profiles of the first frame of each sequence are shown in **b** (with glycerol) and **c** (without glycerol). **d, e**, Localized Ca^{2+} signals produced by single FLIP of NP-EGTA in muscle cells. Image sequences of the Ca^{2+} signal ($\Delta F/F_0$ in pseudo-colour) of a muscle cell without (**d**) and with (**e**) NP-EGTA loading. Boxes indicate the area shown in time sequences, circles illustrate the laser spot, and digits represent milliseconds after a single laser pulse. **f**, Focal increase in $[\text{Ca}^{2+}]_i$ in a *Xenopus* growth cone induced by single FLIP of NP-EGTA. Scale bars, $10 \mu\text{m}$.

of the growth cone was focally elevated by FLIP. Because a relatively long-lasting local Ca^{2+} signal was observed previously during growth cone turning⁵, repetitive FLIP of caged Ca^{2+} was performed. To prevent potential damage to the cells from repetitive laser pulses, we reduced the laser energy to $\sim 0.1 \mu\text{J}$ per pulse. Repetitive laser irradiation every 3 s produced no effect on the morphology and extension of control growth cones over a 30 min period (Fig. 2a). However, repetitive FLIP of NP-EGTA on one side of the growth cone induced a marked turning response of the growth cone towards the side with $[\text{Ca}^{2+}]_i$ elevation within 30 min (Fig. 2b). The attraction involved the preferential protrusion of lamellipodia and filopodia on the FLIP side followed by rapid turning of the growth cone shaft (Fig. 2c; see also Fig. 3a, b). The paths of neurite extension of all the cells loaded with (Fig. 2e) or without (Fig. 2d) NP-EGTA during the 30 min exposure to laser pulses are presented as composite drawings. Of 19 NP-EGTA-loaded growth cones that extended during 30 min of repetitive FLIP, 18 exhibited attractive turning (Fig. 2e), whereas no preferential turning was observed for 14 control growth cones (Fig. 2d). To quantify the turning response, we measured the turning angle and length of neurite extension of each growth cone⁵, where positive angles indicate attraction (Fig. 2f, g). Significant attractive turning was observed for growth cones loaded with NP-EGTA, whereas growth cones without NP-EGTA loading were not affected by the repetitive laser irradiation (average turning angle ($\pm$ s.e.m.): 29.8 ± 4.0 degrees and -5.8 ± 4.8 degrees, respectively; $P < 0.0001$, Kruskal–Wallis test). Both groups of growth cones showed similar lengths of extension over the 30 min period ($12.8 \pm 0.7 \mu\text{m}$ and $9.7 \pm 1.1 \mu\text{m}$, respectively; $P > 0.01$, Kruskal–Wallis test). These results show that direct focal elevation of $[\text{Ca}^{2+}]_i$ on one side of the growth cone is sufficient to specifically initiate attractive turning.

To examine the influence of extracellular Ca^{2+} on the turning behaviour of growth cones induced by the direct focal elevation of $[\text{Ca}^{2+}]_i$, NP-EGTA loaded cells were placed in a Ca^{2+} -free medium¹¹ for turning experiments. The removal or marked decrease in extracellular Ca^{2+} concentration in *Xenopus* cultures increased the rate of neurite extension to about three times that in medium containing millimolar Ca^{2+} (refs 5,12,13). Thus, the turning response of the growth cone was examined after only 10 min of repetitive FLIP for a comparable length of extension to that found in Ca^{2+} -containing medium. The same intensity and frequency of laser pulses were found to cause the growth cone in Ca^{2+} -free medium to turn away from the side of FLIP (repulsion). The switching from attraction to repulsion induced by the same repetitive FLIP in different extracellular media is best illustrated in Fig. 3. The particular growth cone shown was induced to turn four times by repetitive FLIP of NP-EGTA: attraction twice in culture medium and repulsion twice in Ca^{2+} -free medium (summarized in Fig. 3e). By alternating the location of the laser beam between the two sides of the growth cone, it was clear that direct focal elevation of $[\text{Ca}^{2+}]_i$ consistently induced attractive turning in Ca^{2+} -containing culture medium (Fig. 3a, b) and repulsive turning in Ca^{2+} -free medium (Fig. 3c, d). The attractive turning was preceded by the preferential protrusion of lamellipodia and filopodia on the side of FLIP, followed rapidly by the turning of the growth cone shaft. The repulsive turning, in contrast, was accompanied by a local inhibition of filopodia and lamellipodia on the side of FLIP as well as an increased protrusion of lamellipodia and filopodia on the opposite side. Summary and quantification of the turning response for all the growth cones examined in Ca^{2+} -free medium are shown in Fig. 3f and g, respectively. Of 13 NP-EGTA-loaded growth cones, 12 turned away from the FLIP side, a difference that is statistically significant from growth cones without NP-EGTA loading (turning angles -27.2 ± 9.5 degrees and 4.1 ± 6.6 degrees, respectively; $P < 0.005$, Kruskal–Wallis test). Both groups of growth cones extended similar lengths during the 10 min exposure to laser pulses ($13.4 \pm 1.4 \mu\text{m}$ and $14.5 \pm 1.8 \mu\text{m}$, respectively; $P > 0.1$,

Kruskal–Wallis test), indicating the specificity of focal elevation of $[Ca^{2+}]_i$ on the direction of growth cone extension.

The switching from attraction to repulsion does not seem to involve extracellular actions of Ca^{2+} in the medium directly. When

extracellular Ca^{2+} was not completely removed but was decreased to $1.3 \mu M$ using a Ca^{2+} buffer¹⁴, repulsion was still induced by repetitive FLIP of NP-EGTA (9 of 10 growth cones; turning angle -21.1 ± 6.1 degrees; average length $13.5 \pm 1.9 \mu m$). The removal

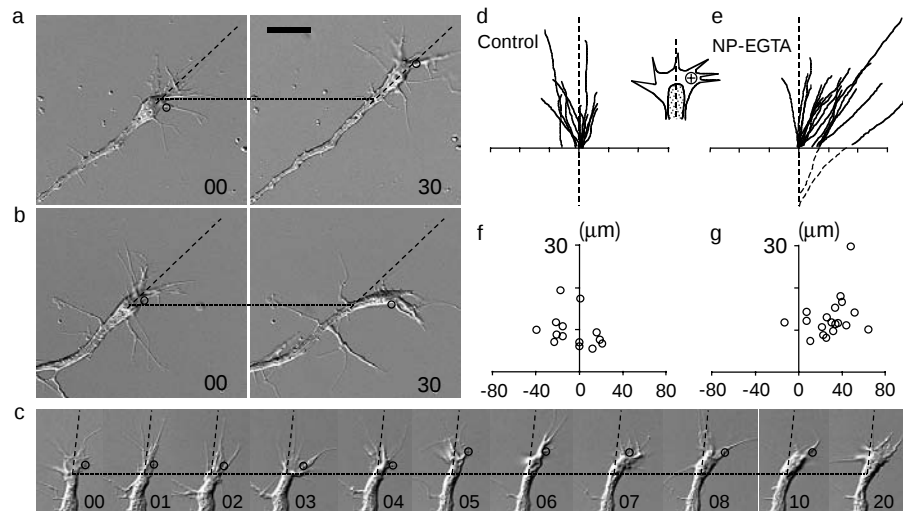


Figure 2 Attractive turning of *Xenopus* growth cones induced by repetitive FLIP of NP-EGTA. **a**, A control growth cone (no NP-EGTA loading) at the beginning and end of 30-min repetitive laser irradiation; circles indicate the laser spot. Dashed lines represent the original direction of extension, and dotted lines represent the corresponding positions along the neurite. **b**, An NP-EGTA-loaded growth cone at the onset and the end of 30-min repetitive FLIP. **c**, A time-lapse sequence of the attractive turning of a growth cone induced by repetitive FLIP of NP-EGTA. Numbers represent minutes after the onset of repetitive FLIP. Scale bars in **a–c**, $20 \mu m$. **d, e**, Composite drawings of the path of growth cones without (**d**) and with (**e**) NP-EGTA loading during a 30-min exposure to FLIP. Each

growth cone was reorientated so that the side of laser irradiation was on the right (illustrated by the diagram in the middle). The origin represents the position of the centre of the growth cone palm at the beginning of the 30-min exposure to the laser. The original direction of growth cone extension was aligned with the vertical axis. Each line depicts the trajectory of the neurite at the end of the 30-min experiment. Tick marks along the horizontal axis represent $5 \mu m$. Note that some growth cones changed their origins, which are represented by the dashed lines. **f, g**, Scatter plots of turning angles plotted against length of neurite extension of all the control (**f**) and NP-EGTA-loaded (**g**) growth cones examined.

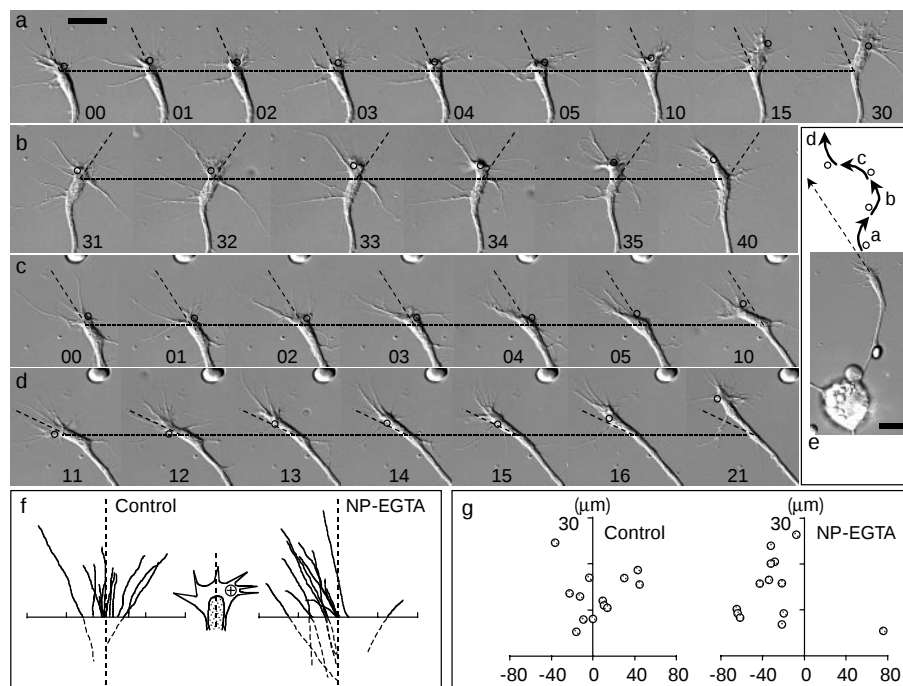


Figure 3 Attractive and repulsive turning of a *Xenopus* growth cone induced by direct focal increase in $[Ca^{2+}]_i$, in different extracellular solutions. **a**, A time-lapse image sequence showing the attractive turning of an NP-EGTA-loaded growth cone induced by repetitive FLIP within 30 min; circles indicate the location of the laser beam. **b**, Attraction was induced for the second time for the same growth cone after repositioning of the laser beam to the other side of the growth cone. **c**, A time-lapse image sequence of the same growth cone, showing repulsive turning when placed in a Ca^{2+} -free medium and exposed to the laser irradiation of same intensity and frequency. **d**, Repulsion was induced for the

second time for the same growth cone when the laser was relocated to the opposite side. Numbers in **a–d** represent minutes after the onset of FLIP. **e**, Schematic summary of the turning responses of the growth cone (twice attractive and twice repulsive); circles indicate the side of the growth cone on which the laser was applied. Scale bars in **a–e**, $20 \mu m$. **f**, Composite drawings of the response of all the growth cones examined in Ca^{2+} -free medium. Tick marks along the horizontal axis represent $5 \mu m$. **g**, Scatter plots of turning angles against length of extension for all the growth cones examined in Ca^{2+} -free medium.

or a significant decrease in extracellular Ca^{2+} concentration is known to affect the resting level of $[\text{Ca}^{2+}]_i$ in *Xenopus* neurons⁵. Fura-2 ratiometric imaging showed that replacing Ca^{2+} -containing culture medium with Ca^{2+} -free solution or low- Ca^{2+} buffer immediately decreased the resting $[\text{Ca}^{2+}]_i$ at the growth cone from 130.3 ± 17.5 nM to 58.5 ± 5.4 nM or 72 ± 5.0 nM, respectively (means $\pm$ s.e.m.; Fig. 4a–c). To determine whether different resting levels of $[\text{Ca}^{2+}]_i$ affect the amount of Ca^{2+} released by FLIP, simultaneous FLIP and fluo-3 imaging were performed. For the reliable detection of small local changes in $[\text{Ca}^{2+}]_i$ induced by FLIP of NP-EGTA, the laser energy was increased to 10 μJ . For seven growth cones examined in culture medium, and subsequently in Ca^{2+} -free medium, similar relative changes in fluo-3 fluorescence ($\Delta F/F_0$, calibrated against the corresponding baseline) at the laser spot were observed upon FLIP (Fig. 4d). This result was confirmed by the average of many growth cones examined independently (Fig. 4e; $P > 0.1$, Student's t -test). For both culture medium and Ca^{2+} -free medium, a single laser pulse induced a $\Delta F/F_0$ of $\sim 60\%$ at the FLIP spot in the first frame of the image sequence. Assuming that fluo-3 fluorescence corresponds linearly to $[\text{Ca}^{2+}]_i$ and by using the resting levels of $[\text{Ca}^{2+}]_i$ measured from fura-2 imaging, the peak levels of $[\text{Ca}^{2+}]_i$ at the FLIP site are estimated to be ~ 78 nM and ~ 35 nM for growth cones in culture medium and in Ca^{2+} -free medium, respectively. It should be emphasized that the Ca^{2+} imaging was performed with laser pulses of ~ 100 times the intensity used for turning experiments. However, according to the theoretical analysis of point source diffusion¹⁵, the concentration of cytosolic Ca^{2+} at the centre of FLIP at the end of the 20-ms charge-coupled-device (CCD) exposure would have decreased to $\sim 1\%$ of that at 1 ms after photolysis during CCD exposure (14 ms for diffusion).

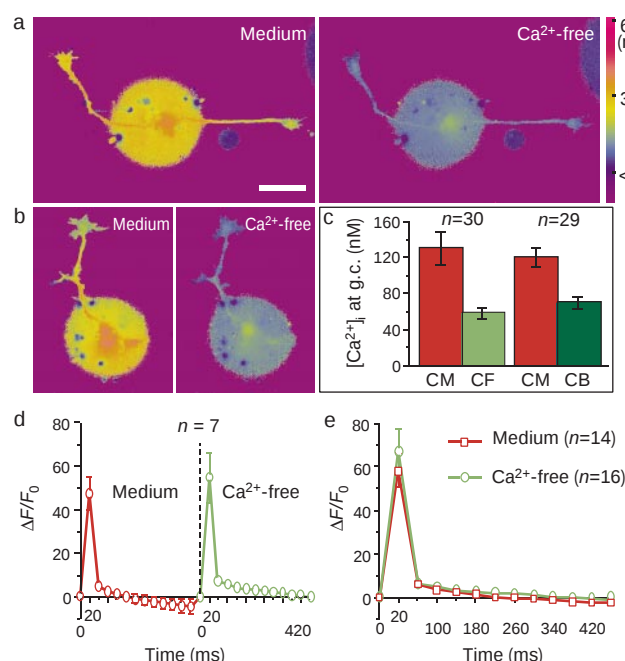


Figure 4 Fluorescence imaging of focal increase in $[\text{Ca}^{2+}]_i$ induced by FLIP of NP-EGTA. **a**, **b**, Fura-2 ratio imaging of the resting $[\text{Ca}^{2+}]_i$ in a bipolar neuron (**a**) and a monopolar neuron (**b**) in culture medium and Ca^{2+} -free medium. The distribution of $[\text{Ca}^{2+}]_i$ is shown in pseudo-colour with the corresponding Ca^{2+} concentration (nM) depicted by the colour bar. **c**, Averages of the resting $[\text{Ca}^{2+}]_i$ at the growth cone from populations of neurons in culture medium (CM) and Ca^{2+} -free medium (CF) or in a Ca^{2+} -buffer containing 1.3 μM free Ca^{2+} (CB). **d**, Fluo-3 imaging of the focal increase in $[\text{Ca}^{2+}]_i$ at the growth cone induced by FLIP of caged Ca^{2+} in culture medium and Ca^{2+} -free medium. No significant difference was observed in the amplitude of focal $[\text{Ca}^{2+}]_i$ increased by FLIP of NP-EGTA for each growth cone examined in these two media. This is confirmed by the average of many growth cones examined (**e**).

Cytosolic buffering of local Ca^{2+} could further decrease the amplitude of the local Ca^{2+} signals that can be detected by the imaging. Taking these two factors into account, the estimate of the actual peak amplitude of the focal $[\text{Ca}^{2+}]_i$ elevated by each laser pulse during growth cone turning would be on the same order as that obtained with our imaging measurement, that is, ~ 78 nM and ~ 35 nM for growth cones in culture medium and Ca^{2+} -free medium, respectively. Therefore, the removal of extracellular Ca^{2+} decreases the resting $[\text{Ca}^{2+}]_i$ as well as the absolute level of focal $[\text{Ca}^{2+}]_i$ elevation induced by FLIP of caged Ca^{2+} with the same laser intensity.

To determine whether the resting $[\text{Ca}^{2+}]_i$ or the absolute level of focal elevation of $[\text{Ca}^{2+}]_i$ contributes to the repulsion, growth cones in Ca^{2+} -free medium were induced to turn by FLIP of NP-EGTA with an increased laser intensity that led to a similar level of focal $[\text{Ca}^{2+}]_i$ as that found in culture medium (~ 70 nM). Of twelve growth cones examined, eight showed marked repulsion, whereas two of the other four growth cones showed no turning and two exhibited attraction (turning angle -15.1 ± 7.5 degrees; average length 17.6 ± 2.4 μm). This result indicates that resting levels of $[\text{Ca}^{2+}]_i$ at the growth cone are important in modulating the turning responses. The opposite turning responses might be a reflection of the differential activation of signalling molecules with different thresholds for $[\text{Ca}^{2+}]_i$: repulsion results from the activation of intracellular molecules with a low threshold for $[\text{Ca}^{2+}]_i$ ($\sim 58 + 35 = 93$ nM), whereas attraction results from the activation of molecules with a high threshold for $[\text{Ca}^{2+}]_i$ ($\sim 130 + 78 = 208$ nM). Under normal $[\text{Ca}^{2+}]_i$ conditions, signalling molecules of the repulsive mechanism might be saturated or desensitized, allowing only attraction after the focal elevation of $[\text{Ca}^{2+}]_i$. Lowering resting $[\text{Ca}^{2+}]_i$ allows the low-threshold molecules to be activated by the local Ca^{2+} signals that reach the threshold for repulsion but remain below the threshold for attraction. The ability to finely manipulate the absolute levels of focal $[\text{Ca}^{2+}]_i$ elevation in the growth cone by using FLIP will permit further experiments to test this hypothesis.

I have shown that direct focal elevation of $[\text{Ca}^{2+}]_i$ on one side of the growth cone is sufficient to initiate both attractive and repulsive turning of the growth cone. Ca^{2+} is a key second messenger that regulates growth cone behaviour^{3,4,16–18}. The guidance of growth cones by many diffusible cues such as netrin-1 has been shown to involve Ca^{2+} (refs 13, 14, 19). However, the precise role of Ca^{2+} signals in directional sensing and steering of the growth cone is unclear. Results from this study provide the first direct evidence that a localized increase in $[\text{Ca}^{2+}]_i$ at the growth cone can provide the directional signal for growth cone extension and is sufficient to steer the growth cones accordingly. Furthermore, the finding that both local and global (resting) cytosolic Ca^{2+} signals contribute to determining the direction of growth cone turning demonstrates an important feature of intracellular signalling in growth cones for processing and integrating various cellular signals. During neural development *in vivo*, the resting $[\text{Ca}^{2+}]_i$ of developing neurons is likely to be influenced and modulated by environmental factors such as adhesion molecules²⁰. Furthermore, raising the cAMP concentration in the growth cone has been shown to decrease high $[\text{Ca}^{2+}]_i$ (ref. 21). Thus, depending on its surroundings, an axon can respond with either attractive or repulsive turning to the same guidance molecules that act through the Ca^{2+} pathway. With recent findings that the turning direction of growth cones in netrin-1 gradients can be regulated by the cyclic AMP pathway (probably downstream of Ca^{2+})¹⁴ as well as by different cytoplasmic components of the appropriate receptors^{22,23}, these results indicate the existence of multiple levels of intracellular regulation of directed growth cone extension. Such diversity of regulation along the signal transduction pathway, if it occurred *in vivo*, could provide the potential for the specific and accurate wiring of millions of axons through a limited number of cues available during development. □

Methods

Optical set up

The laser beam (wavelength 337 nm) from a pulsed nitrogen laser (model no. VSL-337N, Laser Science) was introduced into a Nikon TE300 inverted microscope with a laser-to-microscope adapter (Laser Science) mounted on a modified fluorescence epi-illuminator. A dichroic mirror (DM400, Nikon) in the epi-illuminator reflected the laser beam into the microscope while allowing simultaneous fluorescence imaging by using excitation light with wavelengths >400 nm (thus excluding fura-2 imaging). The laser beam was focused into a spot of ~ 2 μ m in diameter with a 40 $\times$ /1.3 oil-immersion objective (Plan-Fluor., Nikon). A frame-transfer cooled CCD camera (PXL37, Photometrics) was used with the AIW software from Axon Instruments for image acquisition. Each image was acquired by a 20 ms exposure with CCD binning and gain set at 4 and 8, respectively. A sample rate of 25 Hz for a small region of 50×50 (pixel)² was normally achieved. The FLIP and imaging were synchronized by a trigger signal from the computer to fire the laser 5 ms after the onset of the first CCD exposure of the image sequence.

Growth cone turning

Xenopus neurons from 6-h cultures^{24,25} were incubated with 6 μ M acetoxymethyl (AM) ester derivative of NP-EGTA (Molecular Probes) for 30 min, followed by three washes. Growth cones frequently exhibited a blunt morphology with little motility immediately after loading, so they were left to recover in culture medium for 3 h. This 3-h incubation also allowed further de-esterification of NP-EGTA AM and Ca²⁺ binding by NP-EGTA. Turning experiments were performed in an open microscopy chamber²⁶. Each growth cone was positioned such that the laser spot would be located on one side and be kept at the same location throughout the entire experiment. Images of the growth cone at various times after the onset of repetitive FLIP were processed digitally and acquired²⁵, and quantitative measurement and analysis of the turning response were then performed on these digital images^{5,11}.

Ca²⁺ imaging

Cells were loaded with 8 μ M fluo-3 AM (Molecular Probes) 2 h after the end of NP-EGTA loading for 30 min followed by a 30 min incubation in culture medium. The Ca²⁺ signals were measured by using the change in fluorescence relative to the baseline before the irradiation ($\Delta F/F_0$). All the fluo-3 fluorescence was corrected for background by using the fluorescence from the same cell lysed at the end of the experiment with digitonin²⁷. Focal increase in [Ca²⁺]_i was measured at the FLIP site with a 2- μ m diameter zone. For visual presentation of the Ca²⁺ signals, the resulting $\Delta F/F_0$ was given a threshold of 15%, pseudo-coloured and overlaid on a grey-scale image of the cell before the laser irradiation. Fura-2 measurements of the resting [Ca²⁺]_i was performed with previous methods^{5,28}. The cells were typically imaged first in culture medium and then in Ca²⁺-free medium.

Received 14 September; accepted 20 October 1999.

- Bray, D. & Hollenbeck, P. J. Growth cone motility and guidance. *Annu. Rev. Cell Biol.* **4**, 43–61 (1988).
- Tessier-Lavigne, M. & Goodman, C. S. The molecular biology of axon guidance. *Science* **274**, 1123–1133 (1996).
- Kater, S. B. & Mills, L. R. Regulation of growth cone behavior by calcium. *J. Neurosci.* **11**, 891–899 (1991).
- Gu, X. & Spitzer, N. C. Distinct aspects of neuronal differentiation encoded by frequency of spontaneous Ca²⁺ transients. *Nature* **375**, 784–787 (1995).
- Zheng, J. Q., Felder, M., Connor, J. A. & Poo, M. M. Turning of nerve growth cones induced by neurotransmitters. *Nature* **368**, 140–144 (1994).
- Zheng, J. Q., Poo, M. M. & Connor, J. A. Calcium and chemotropic turning of nerve growth cones. *Perspect. Dev. Neurobiol.* **4**, 205–213 (1996).
- Ishihara, A., Gee, K., Schwartz, S., Jacobson, K. & Lee, J. Photoactivation of caged compounds in single living cells: an application to the study of cell locomotion. *Biotechniques* **23**, 268–274 (1997).
- Ellis-Davies, G. C. & Kaplan, J. H. Nitrophenyl-EGTA, a photolabile chelator that selectively binds Ca²⁺ with high affinity and releases it rapidly upon photolysis. *Proc. Natl Acad. Sci. USA* **91**, 187–191 (1994).
- Connor, J. A. & Ahmed, Z. Diffusion of ions and indicator dyes in neural cytoplasm. *Cell. Mol. Neurobiol.* **4**, 53–66 (1984).
- Ahmed, Z. & Connor, J. A. Calcium regulation by and buffer capacity of molluscan neurons during calcium transients. *Cell. Calcium* **9**, 57–69 (1988).
- Zheng, J. Q., Wan, J. J. & Poo, M. M. Essential role of filopodia in chemotropic turning of nerve growth cone induced by a glutamate gradient. *J. Neurosci.* **16**, 1140–1149 (1996).
- Bixby, J. L. & Spitzer, N. C. Early differentiation of vertebrate spinal neurons in the absence of voltage-dependent Ca²⁺ and Na⁺ influx. *Dev. Biol.* **106**, 89–96 (1984).
- Song, H. J., Ming, G. L. & Poo, M. M. cAMP-induced switching in turning direction of nerve growth cones. *Nature* **388**, 275–279 (1997).
- Ming, G. L. *et al.* cAMP-dependent growth cone guidance by netrin-1. *Neuron* **19**, 1225–1235 (1997).
- Lohof, A. M., Quillan, M., Dan, Y. & Poo, M. M. Asymmetric modulation of cytosolic cAMP activity induces growth cone turning. *J. Neurosci.* **12**, 1253–1261 (1992).
- Kater, S. B. & Rehder, V. The sensory-motor role of growth cone filopodia. *Curr. Opin. Neurobiol.* **5**, 68–74 (1995).
- Gundersen, R. W. & Barrett, J. N. Characterization of the turning response of dorsal root neurites toward nerve growth factor. *J. Cell Biol.* **87**, 546–554 (1980).
- Gomez, T. M. & Spitzer, N. C. *In vivo* regulation of axon extension and pathfinding by growth-cone calcium transients. *Nature* **397**, 350–355 (1998).
- Song, H. *et al.* Conversion of neuronal growth cone responses from repulsion to attraction by cyclic nucleotides. *Science* **281**, 1515–1518 (1998).
- Snow, D. M., Atkinson, P. B., Hassinger, T. D., Letourneau, P. C. & Kater, S. B. Chondroitin sulfate proteoglycan elevates cytoplasmic calcium in DRG neurons. *Dev. Biol.* **166**, 87–100 (1994).

- Lankford, K. L. & Letourneau, P. C. Roles of actin filaments and three second-messenger systems in short-term regulation of chick dorsal root ganglion neurite outgrowth. *Cell Motil. Cytoskel.* **20**, 7–29 (1991).
- Hong, K. *et al.* A ligand-gated association between cytoplasmic domains of UNC5 and DCC family receptors converts netrin-induced growth cone attraction to repulsion. *Neuron* **97**, 427–441 (1999).
- Bashaw, G. J. & Goodman, C. S. Chimeric axon guidance receptors: the cytoplasmic domains of slit and netrin receptors specify attraction versus repulsion. *Neuron* **97**, 917–926 (1999).
- Tabti, N. & Poo, M. M. in *Culturing Nerve Cells* (ed. Banker, G. & Goslin, K.) 137–154 (MIT, Cambridge, Massachusetts, 1990).
- Wang, Q. & Zheng, J. Q. Cyclic AMP-mediated regulation of neurotrophin-induced collapse of nerve growth cones. *J. Neurosci.* **18**, 4973–4984 (1998).
- Ming, G. L., Lohof, A. M. & Zheng, J. Q. Acute morphogenic and chemotropic effects of neurotrophins on cultured embryonic *Xenopus* spinal neurons. *J. Neurosci.* **17**, 7860–7871 (1997).
- Kao, J. P., Harootyan, A. T. & Tsien, R. Y. Photochemically generated cytosolic calcium pulses and their detection by fluo-3. *J. Biol. Chem.* **264**, 8179–8184 (1989).
- Tsien, R. Y. Fluorescence measurement and photochemical manipulation of cytosolic free calcium. *Trends Neurosci.* **11**, 419–424 (1988).

Acknowledgements

I thank R. Maki Fitzsimonds (Yale University, New Haven, Connecticut, USA) and J. Alder (University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, Piscataway, New Jersey, USA) for their critical review and comments on the manuscript, and J. Gibney for technical support. The initial part of this study was performed at the Marine Biological Laboratory (Woods Hole, Massachusetts, USA) with the support of a Nikon summer fellowship. This work was supported by a grant from the National Institutes of Health.

Correspondence and requests for materials should be addressed to J.Q.Z. (e-mail: zhengjq@umdj.edu).

Calcium signalling in the guidance of nerve growth by netrin-1

Kyonsoo Hong[†], Makoto Nishiyama[†], John Henley*, Marc Tessier-Lavigne[‡] & Mu-ming Poo*

* Department of Biology, University of California at San Diego, La Jolla, California 92093-0357, USA

‡ Howard Hughes Medical Institute, Department of Anatomy and Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, California 94143-0452, USA

† These authors contributed equally to this work

Pathfinding by growing axons in the developing nervous system is guided by diffusible or bound factors that attract or repel the axonal growth cone^{1,2}. The cytoplasmic signalling mechanisms that trigger the responses of the growth cone to guidance factors are mostly unknown³. Previous studies have shown that the level and temporal patterns of cytoplasmic Ca²⁺ can regulate the rate of growth-cone extension *in vitro*^{4–8} and *in vivo*⁹. Here we report that Ca²⁺ also mediates the turning behaviour of the growth cones of cultured *Xenopus* neurons that are induced by an extracellular gradient of netrin-1, an established diffusible guidance factor *in vivo*^{1,10}. The netrin-1-induced turning response depends on Ca²⁺ influx through plasma membrane Ca²⁺ channels, as well as Ca²⁺-induced Ca²⁺ release from cytoplasmic stores¹¹. Reduction of Ca²⁺ signals by blocking either of these two Ca²⁺ sources converted the netrin-1-induced response from attraction to repulsion. Activation of Ca²⁺-induced Ca²⁺ release from internal stores with a gradient of ryanodine in the absence of netrin-1 was sufficient to trigger either attractive or repulsive responses, depending on the ryanodine concentration used. These results support the model that cytoplasmic Ca²⁺ signals mediate growth-cone guidance by netrin-1, and different patterns of Ca²⁺ elevation trigger attractive and repulsive turning responses.

We created a microscopic gradient of netrin-1 in cell cultures by repetitive pulsatile application of netrin-1 from a micropipette¹².

When the gradient was applied near the growth cones of cultured *Xenopus* spinal neurons, we observed marked attractive turning responses towards the source of netrin-1¹³ (Fig. 1a, g, h). Reduction of the extracellular Ca^{2+} concentration from 1 mM to 1 μM abolished both attractive and repulsive turning responses induced by netrin-1¹³, suggesting that Ca^{2+} signalling is involved in the turning response. We found that Ca^{2+} influx is required for netrin-1-induced attraction, as the attraction was abolished by blocking plasma membrane Ca^{2+} channels with Cd^{2+} (50 μM , Fig. 1b, g, h). Calcium release from internal stores also appeared to be necessary, as depletion of Ca^{2+} stores by pre-incubation of the cells with benzyhydroquinone¹⁴ (BHQ, 10 μM) or thapsigargin¹⁵ (2 μM) blocked the netrin-1-induced attraction (Fig. 1c, g, h). When nimodipine (10 μM), a specific L-type Ca^{2+} -channel blocker¹⁶, was added to the culture medium, however, the same netrin-1 gradient induced marked repulsion of the growth cones (Fig. 1d, g, h). The same repulsion was observed when ryanodine was added to the bath at a concentration (100 μM) that is known to trap the ryanodine receptor in the closed state¹⁷ (Fig. 1e, g, h). Finally, blockade of both L-type channels and Ca^{2+} -induced Ca^{2+} release

(CICR) with nimodipine and ryanodine completely abolished the turning response of the growth cone (Fig. 1f, g, h). In the presence of BHQ, thapsigargin and ryanodine, the net extension of the growth cone during the 1-h period of the turning assay was reduced, but sufficient growth took place to allow measurements of the turning angle. Together, these results are consistent with netrin-1-induced turning being mediated by cytoplasmic Ca^{2+} : high-level Ca^{2+} signals in the normal medium triggered the attractive response, whereas low-level Ca^{2+} signals, resulting from blocking either Ca^{2+} influx through L-type channels or CICR, led to the repulsive response. Blocking both L-type channels and CICR resulted in a complete absence of netrin-1-dependent Ca^{2+} signals, so there was no preferential turning response.

To monitor the Ca^{2+} level in the growth cone, we microinjected a fluorescent Ca^{2+} -sensitive dye, Oregon Green BAPTA-dextran¹⁸, into individual neurons. A significant increase in the average Ca^{2+} level was observed at the growth cone close to the source of netrin-1, but not at a more distant growth cone of the same neuron (Fig. 2a), within minutes after the onset of the netrin-1 gradient. In several large growth cones, we also observed the transient appearance of

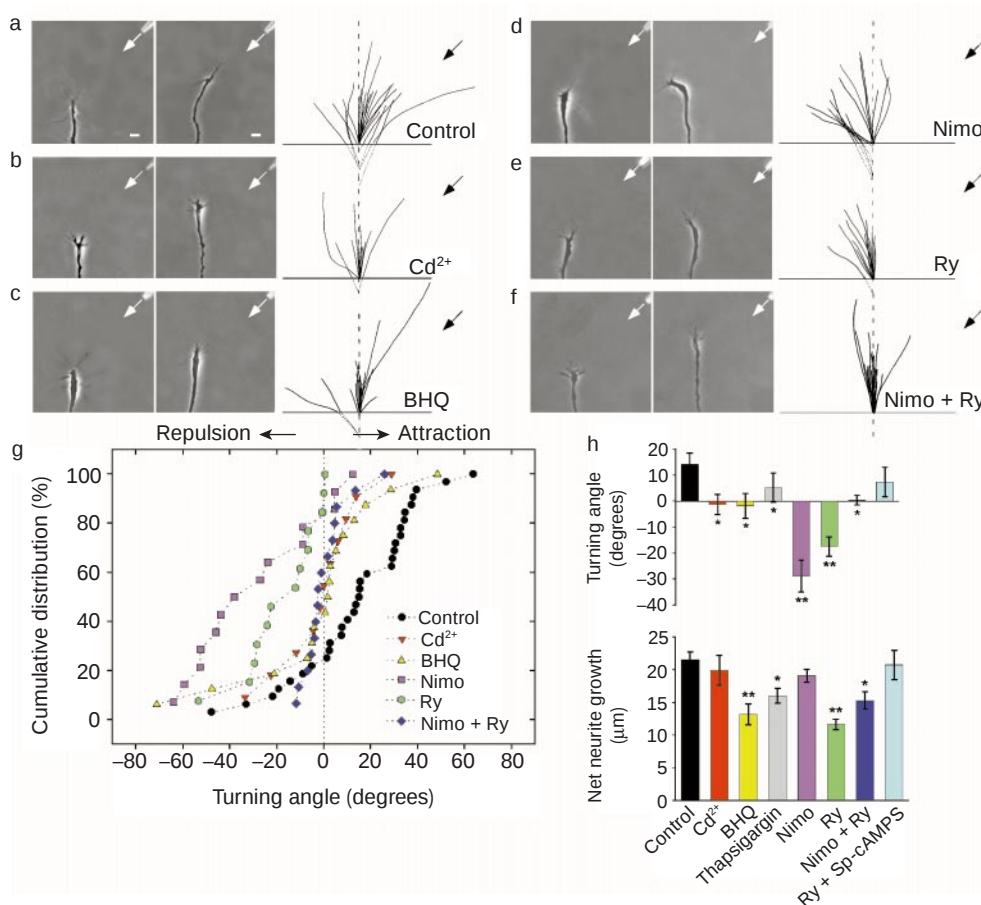


Figure 1 Turning of growth cones of *Xenopus* spinal neurons induced by a gradient of netrin-1. **a**, A gradient of netrin-1 was applied to the growth cone by pulsatile application of picolitres of netrin-1 solution (5 $\mu\text{g ml}^{-1}$) from a micropipette (see Methods). Micrographs were recorded at the onset (left image) and the end (right image) of a 1-h exposure to a netrin-1 gradient. Scale bar, 10 μm . Superimposed traces on the right depict the trajectory of neurite extension during the 1-h period for all the neurons examined. The origin is the centre of the growth cone at the onset of the experiment, and the original direction of growth was vertical. Dashed traces depict cases in which the neurite shank shifted during the assay. The arrow indicates the direction of the gradient. **b–f**, Experiments carried out in the same manner as in **a**, except that CdCl_2 (Cd^{2+} , 50 μM), BHQ (10 μM), nimodipine (Nimo, 10 μM), ryanodine (Ry, 100 μM) and nimodipine (10 μM) + ryanodine (100 μM), respectively, were added to the culture

medium. **g**, The distribution of turning angles. For each experimental condition, the angular positions of all growth cones at the end of a 1-h exposure to a netrin-1 gradient are shown in the cumulative frequency plot. The percentage value refers to the percentage of growth cones with angular position $\leq$ a given angular value. Data shown are turning observed for neurons in normal culture medium and medium containing various drugs as shown in **b–f**. **h**, The average turning angle and the average neurite extension during the 1-h assay for experiments shown in **a–f** as well as experiments using thapsigargin and ryanodine + Sp-cAMPS. Results that show significant differences from the control group (exposed to a netrin-1 gradient in the absence of the drug treatment) are marked: asterisk, $P < 0.05$; double asterisks, $P < 0.001$; Mann-Whitney rank sum test.

gradients of Ca^{2+} , with the level higher on the side facing the netrin-1 source (Fig. 2b). These transient Ca^{2+} gradients dissipated quickly across the growth cone, resulting in a global elevation of cytosolic Ca^{2+} . It is possible that the Oregon Green BAPTA-dextran had facilitated the dissipation of cytosolic Ca^{2+} gradients by serving as a mobile buffer, thus destroying more pronounced Ca^{2+} gradients that could have occurred naturally across the growth cone.

The degree of netrin-1-induced changes in Ca^{2+} over the entire growth cone varied greatly among the growth cones examined; a significant elevation ($\Delta F/F > 5\%$) occurred 10–15 min after the onset of the netrin-1 gradient in 20 out of 30 neurons (Fig. 2c). In many cases, the Ca^{2+} elevation was relatively sustained in the presence of the netrin-1 gradient (Fig. 2a, e) but dissipated to the control level after the gradient was terminated, as shown by experiments in which the netrin-1 gradient was repeatedly applied to the same growth cone (Fig. 2d). Furthermore, the Ca^{2+} elevation was completely absent when netrin-1 was applied in the presence of Cd^{2+} (Fig. 2e), consistent with the notion that Ca^{2+} influx is required for triggering CICR in response to netrin-1. The average magnitude of Ca^{2+} elevation over the entire growth cone was similarly reduced in the presence of nimodipine, ryanodine, BHQ or thapsigargin (Fig. 2e; and data not shown). However, the effects of nimodipine/ryanodine on the turning response were different from those of BHQ/thapsigargin (repulsion compared with no turning). This suggests that downstream effects of the Ca^{2+} signal may also depend on the type of Ca^{2+} source. Depletion of all internal stores by BHQ/thapsigargin appeared to be sufficient to abolish all turning responses, despite some Ca^{2+} elevation due to Ca^{2+} influx.

An optimal range of cytosolic Ca^{2+} levels is required for normal

growth-cone motility⁵. The cytosolic Ca^{2+} elevation that we observed may provide merely a permissive condition for the growth-cone turning induced by netrin-1. Alternatively, it may act as a signal to trigger the turning response. If so, a gradient of cytosolic Ca^{2+} should be sufficient to trigger the turning response in the absence of the netrin-1 gradient. Indeed, when a gradient of ryanodine (1 μM in the pipette) was created across the growth cone, we observed marked attraction towards the source of ryanodine (Fig. 3a, e, f). Under this experimental condition, the extracellular concentration of ryanodine at the growth cone was estimated to be in the nanomolar range¹² (see Methods), a concentration that is known to be effective in activating CICR from internal stores¹⁷. Fluorescence imaging of Ca^{2+} with Oregon Green BAPTA-dextran showed an elevation of Ca^{2+} in the growth cone (in 7 out of 10 neurons) within minutes after the onset of this 'low-dose' ryanodine gradient (Fig. 4a, c). Similar to that for the netrin-1 gradient, we detected transient gradients of Ca^{2+} across the growth cone only in some cases, with higher Ca^{2+} on the side facing the ryanodine pipette. Thus, a gradient of CICR in the cytoplasm was apparently responsible for triggering the attractive response. Interestingly, when the concentration of ryanodine in the pipette was increased to 1 mM (1 μM extracellularly at the growth cone), growth cones exhibited marked repulsive turning away from the source of ryanodine (Fig. 3b, e, f). At this 'high-dose' of ryanodine, Ca^{2+} -release channels associated with CICR are largely inhibited¹⁷ and thus may result in a shallower gradient of Ca^{2+} release from internal stores. This is consistent with the observation made with Oregon Green BAPTA-dextran, which showed no overall increase in the Ca^{2+} level after exposure to the high-dose ryanodine gradient

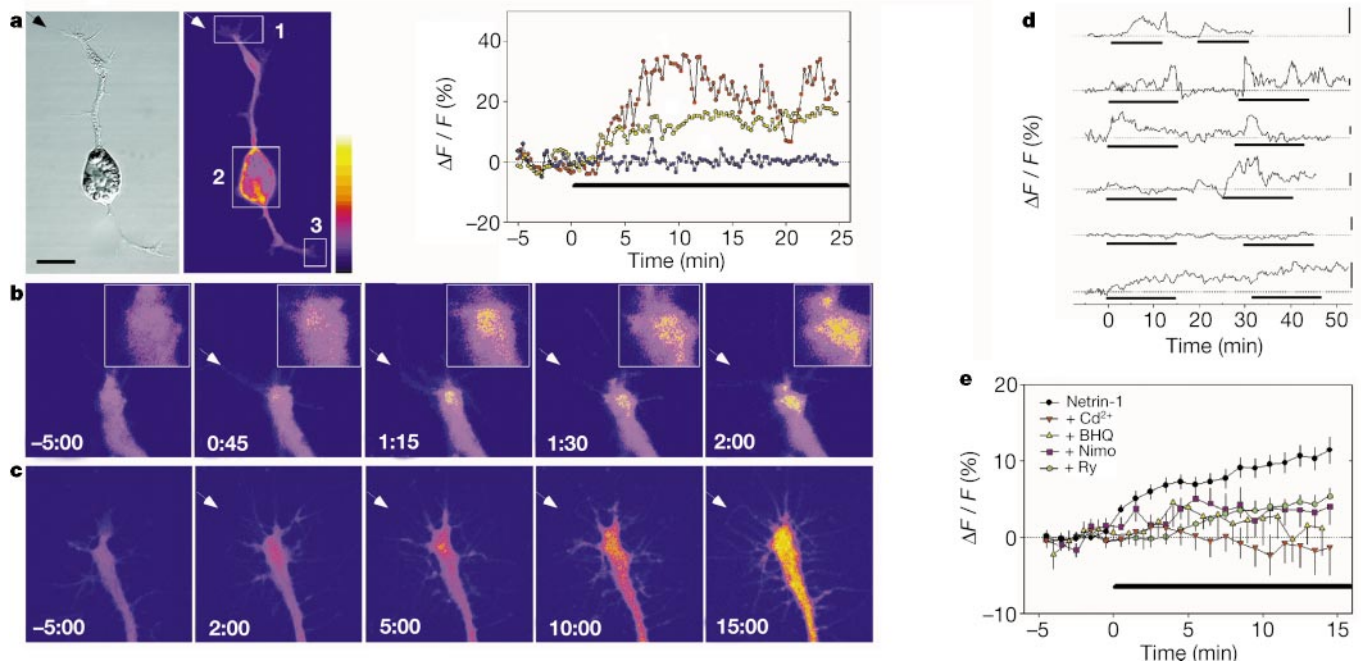


Figure 2 Calcium elevation induced by a gradient of netrin-1. **a**, Bright-field and fluorescence images of a *Xenopus* spinal neuron, which was exposed to a gradient of netrin-1. The neuron was microinjected with Oregon Green BAPTA-dextran and the pipette delivering the netrin-1 (5 $\mu\text{g ml}^{-1}$ in the pipette) was located 100 μm from the growth cone on the top (arrow). Scale bar, 20 μm . (The intensity of the fluorescence is coded by pseudo-colours in linear scale, as shown on the right, with white and yellow representing the highest intensities.) Traces on the right depict changes in the total fluorescence at the growth cones (boxes 1 and 3) and at the soma (box 2) with time before and after the onset of the netrin-1 gradient (indicated by the bar). **b**, **c**, Images of Oregon Green BAPTA-dextran fluorescence of two growth cones at different times (min:s) before and after the onset of the netrin-1 gradient. Note the transient cytoplasmic gradient of fluorescence in **b**, as shown in higher magnification in the insets. **d**, Normalized changes

in Ca^{2+} levels demonstrated by percentage changes in Oregon Green BAPTA-dextran fluorescence ($\Delta F/F$) with time, as observed in six sample growth cones after two episodes of exposure to the netrin-1 gradient (horizontal bars). The fluorescence was normalized to the average value observed before the onset of the first netrin-1 gradient. Scale bar, 10%. **e**, Summary of percentage change in fluorescence induced by a netrin-1 gradient in the normal culture medium (filled circles, $n = 30$) and in medium supplemented with nimodipine (25 μM , filled squares, $n = 10$), ryanodine (150 μM , filled circles, $n = 10$), CdCl_2 (50 μM , inverted triangles, $n = 10$) or BHQ (10 μM , filled triangles, $n = 10$). The value of fluorescence from each growth cone was normalized to the mean value observed during the period before the onset of the gradient before averaging. The error bars represent s.e.m.

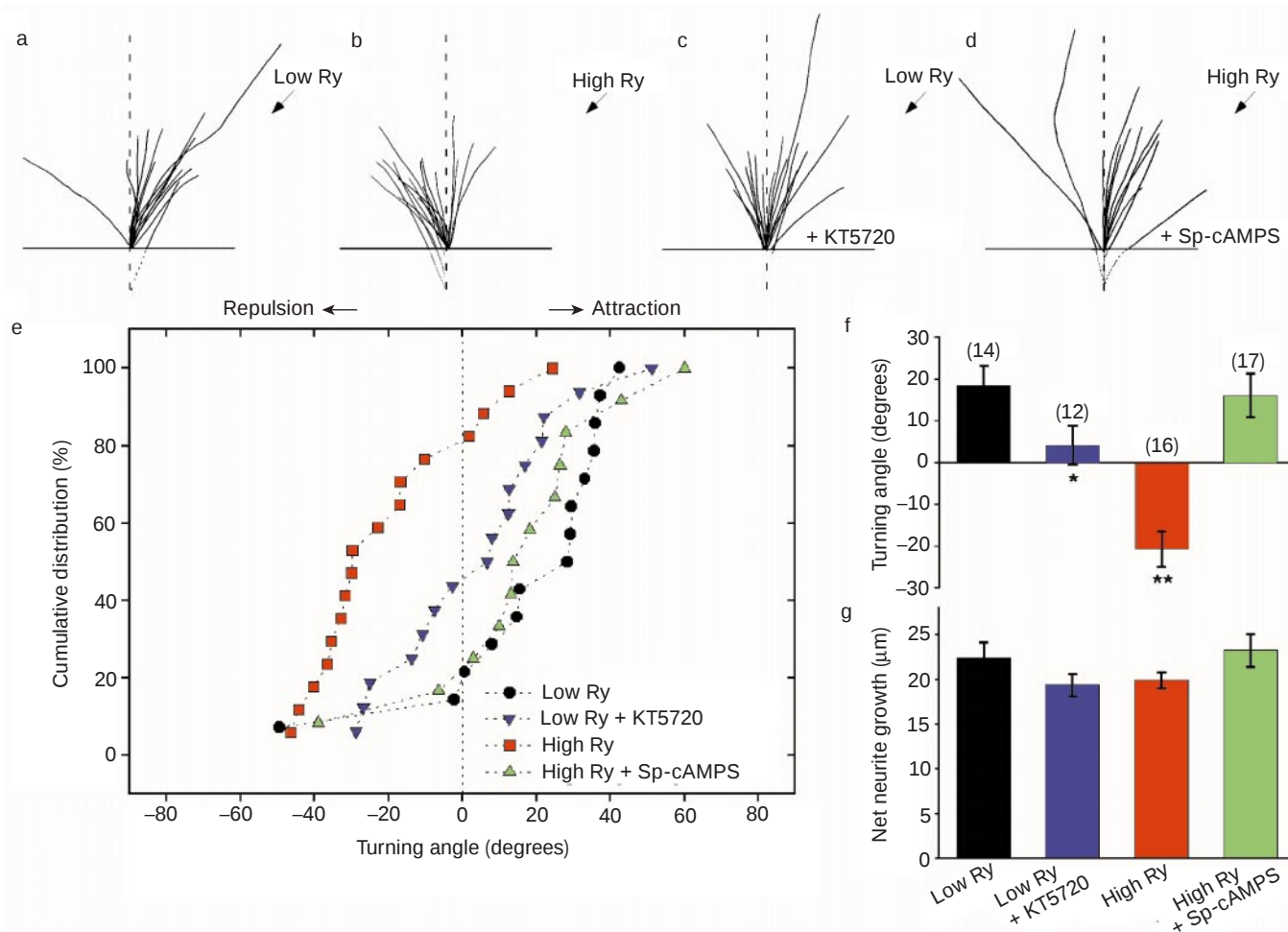


Figure 3 Turning responses induced by a gradient of ryanodine. **a–d**, Trajectories of neurite extension after 1-h exposure to a ryanodine gradient, with low (1 μ M) (**a**, **c**) and high (1 mM) (**b**, **d**) concentrations of ryanodine in the pipette. In **c** and **d**, KT5720 (200 nM) and Sp-cAMPS (20 μ M) were added to the bath, respectively. **e**, Distribution of turning

angles. Data are from the same set of experiments as in **a–d**. **f**, The average turning angle and net neurite extension during the 1-h assay for experiments shown in **a–e**. Results with significant differences from the control group (low ryanodine) are marked: asterisk, $P < 0.05$; double asterisk, $P < 0.001$; Mann-Whitney rank sum test.

(Fig. 4b, d). Instead, a slight reduction in the overall Ca^{2+} level was observed, although transient low-level gradients of Ca^{2+} may have occurred. Finally, we found that at an intermediate concentration of ryanodine (100 μ M in the pipette), there were diverse turning responses, either attractive or repulsive, with an average turning angle near zero ($-2.5 \pm 5.7^\circ$, $n = 10$). Together, these experiments with ryanodine gradients indicate that Ca^{2+} release from internal stores is sufficient to trigger either attractive or repulsive responses, depending on the magnitude and/or pattern of Ca^{2+} elevation in the growth cone. Our findings are also consistent with attractive turning induced by a gradient of acetylcholine, which triggers a gradient of Ca^{2+} elevation in these growth cones¹⁹, and with the induction of filopodial activity²⁰ and the turning of growth cones²¹ triggered by localized photolytic release of Ca^{2+} from caged Ca^{2+} compounds.

Previous studies^{13,22,23} have shown that the cAMP level in the neuron is a critical determinant of the turning response of these *Xenopus* growth cones. Attraction of the growth cone towards the netrin-1 source is converted to repulsion when cAMP-dependent signalling pathways are inhibited by Rp-cAMPS, a membrane-permeable competitive analogue of cAMP, or by KT5720, an inhibitor of protein kinase A (PKA)¹³. As *in situ* activation of ryanodine receptors requires cAMP-dependent phosphorylation²⁴, cAMP may be modifying CICR. We found that bath application of KT5720 resulted in a significant reduction in the attractive response induced by a low-dose ryanodine gradient (Fig. 3c, e, f), whereas similar treatment had no effect on the repulsion induced by a high-

dose gradient of ryanodine (turning angle $-20 \pm 6^\circ$, $n = 3$). Conversely, treatment with Sp-cAMPS (20 μ M), a membrane-permeable analogue of cAMP, converted the turning induced by a high-dose ryanodine gradient from repulsion to attraction (Fig. 3d, e, f). Sp-cAMPS was also effective in converting the response induced by the netrin-1 gradient in the presence of ryanodine (100 μ M in the bath) from repulsion to attraction (Fig. 1h). Thus modulation of CICR by PKA may be responsible for cAMP-dependent switching of growth-cone turning responses in a netrin-1 gradient¹³. In support of this idea, fluorescence imaging of Ca^{2+} showed that Ca^{2+} elevation induced by the low-dose ryanodine gradient was reduced by PKA inhibition with KT5720 (Fig. 4c), whereas Sp-cAMPS markedly enhanced the Ca^{2+} signal induced by the high-dose ryanodine gradient (Fig. 4d).

Our results support the following model of the Ca^{2+} signalling that underlies growth cone guidance by netrin-1. An extracellular gradient of netrin-1 induces a gradient of activation of the netrin-1 receptor DCC (deleted in colorectal cancer)¹⁰ in the plasma membrane. Downstream events in the cytoplasm, possibly mediated by phospholipase C- γ and phosphoinositide-3-kinase²⁵, result in a gradient of activation of plasma membrane Ca^{2+} channels²⁶, including L-type channels. Increased Ca^{2+} influx through Ca^{2+} channels in turn triggers a gradient of CICR from internal stores, resulting in a cytosolic gradient of Ca^{2+} across the growth cone. A steep gradient of Ca^{2+} in the growth cone may activate a gradient of cytoskeletal polymerization, leading to the attractive turning, whereas a shallow

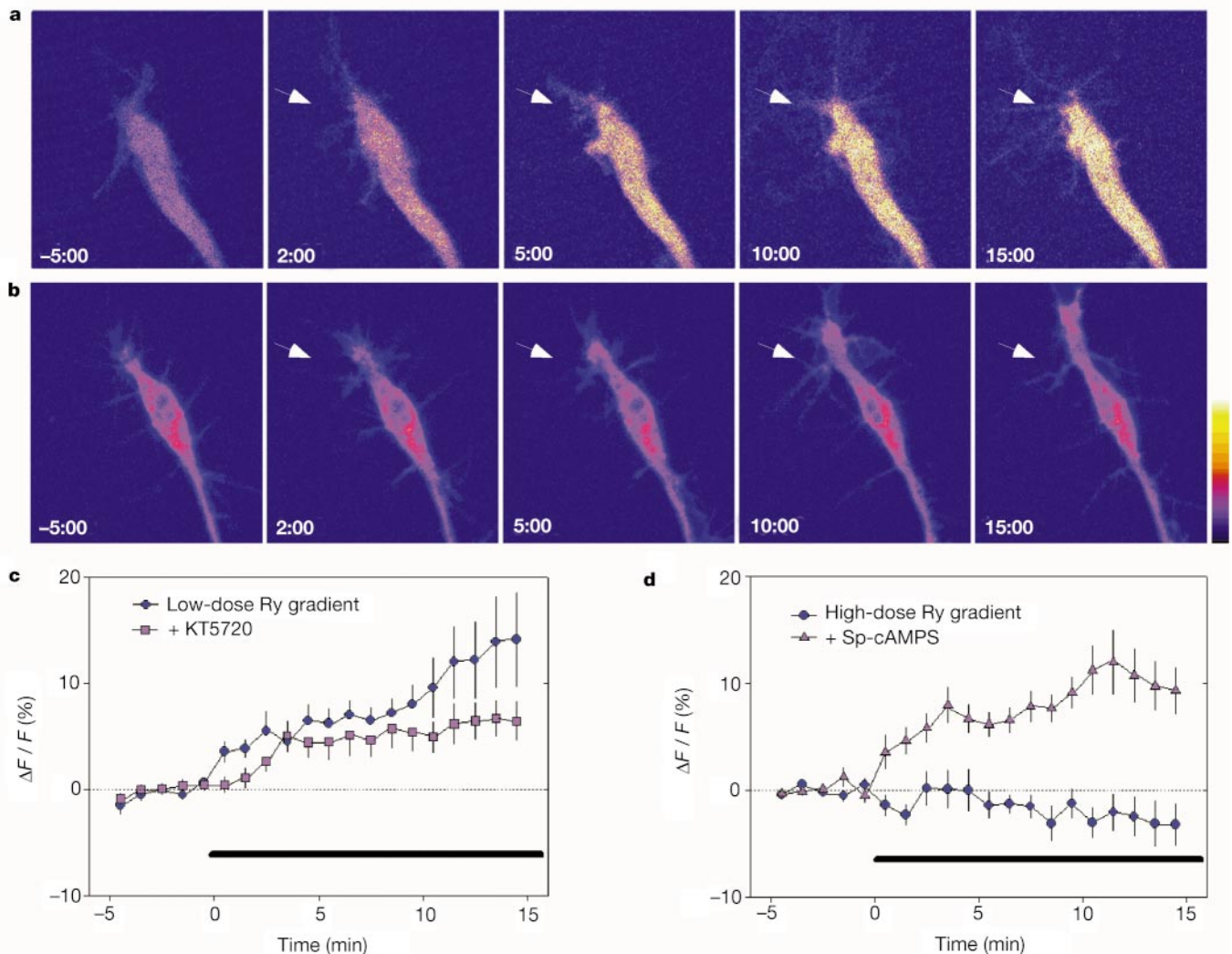


Figure 4 Calcium signals induced by gradients of ryanodine. **a, b**, Images of Oregon Green BAPTA-dextran fluorescence (see Fig. 2a) recorded from two growth cones at different times (min:s) before and after the onset of a ryanodine gradient. The ryanodine concentration in the pipette was 1 μ M and 1 mM in **a** and **b**, respectively. Note the increase of fluorescence in **a** but not in **b**. **c, d**, Summary of fluorescence changes

induced by a low-dose ryanodine gradient (1 μ M in the pipette, $n = 10$) (**c**) or high-dose ryanodine gradient (1 mM in the pipette, $n = 12$) (**d**). Fluorescence changes induced by a low-dose ryanodine gradient in the presence of KT5720 (400 nM, $n = 14$) (**c**) and by high-dose ryanodine gradient in the presence of bath-applied Sp-cAMPS (20 μ M, $n = 12$) (**d**) are also shown.

gradient of Ca^{2+} may activate a gradient of depolymerization, leading to the repulsive response. In the CA1 region of the hippocampus, Ca^{2+} -dependent kinases and phosphatases may be differentially activated by different amplitudes or patterns of Ca^{2+} elevation, resulting in long-term synaptic potentiation and depression^{27–29}, respectively. In an analogous manner, differential activation of enzymes by Ca^{2+} in the growth cone may be responsible for triggering polymerization or depolymerization of cytoskeletal elements, leading to either attractive or repulsive turning responses.

We have shown that Ca^{2+} influx and CICR are involved in mediating growth-cone guidance induced by the guidance molecule netrin-1. Our results point to the importance of the precise profile of Ca^{2+} gradients in determining the turning response of the growth cone. Diametrically opposite turning behaviours can be triggered by different patterns of CICR, which underscores the dynamic and versatile nature of Ca^{2+} signalling in the cytoplasm. □

Methods

Culture preparation

Cultures of spinal neurons were prepared from the neural tube tissue of 1-day-old *Xenopus* embryos as described³⁰. The cultures were used between 14 and 20 h after plating at room temperature (20–22 °C). The culture medium consisted of 49% (v/v) Leibovitz medium

(GIBCO), 1% (v/v) fetal bovine serum (HyClone) and 50% (v/v) Ringer's solution (in mM): 115 NaCl, 2 CaCl_2 , 2.5 KCl and 10 HEPES pH 7.4. The experiments were carried out at room temperature in culture medium. Chemicals were added to the culture medium for at least 30 min for pharmacological treatments and were present during the experiments.

Production of chemical gradients and growth-cone turning assay

Microscopic gradients of chemicals were produced as described¹². Briefly, repetitive pressure injection of picolitre volumes of solutions containing the guidance cue was applied through a micropipette with a tip opening of $\sim 1 \mu\text{m}$. The pressure was applied with an electrically gated pressure application system (Picospritzer, General Valve). A standard pressure pulse of 3 p.s.i. in amplitude and 20 ms in duration was applied to the pipette at a frequency of 2 Hz, using a pulse generator (SD9, Grass Instruments). Theoretical analysis¹² and direct measurements of the gradient using fluorescent dyes³⁰ have shown that the average concentration of the chemical is about 10-fold lower at the growth cone than in the pipette. The concentration gradient across the growth cone (typical width of 10 μm) at a distance of 100 μm from the pipette tip was 5–10%. Different concentrations of the chemicals in the pipette result in different average extracellular concentrations of the chemical at the growth cone, but the same gradient. The total volume of the culture medium in the culture dish during the experiment was 4 ml. The final concentration of factors ejected from the pipette in the bath at the end of the 1-h experiments was estimated to be $\sim 10^5$ lower than that in the pipette. Microscopic images of neurites were recorded with a CCD camera (Toshiba IK-541RA) attached to a phase contrast microscope (Nikon TMS) and stored in a microcomputer for later analysis using NIH Image. To determine the total length of neurite extension, the entire trajectory of the neurite at the end of the 1-h period was measured with a digitizer. To assay growth-cone turning, the pipette tip was placed 100 μm away from the centre of the growth cone and at

an angle of 45° with respect to the initial direction of neurite extension (indicated by the last 10-µm segment of the neurite). The turning angle was defined by the angle between the original direction of neurite extension and a straight line connecting the positions of the growth cone at the onset and the end of the 1-h period. Only those growth cones with net extension >7 µm over the 1-h period were included for analysis.

Calcium imaging

Isolated *Xenopus* spinal neurons were microinjected at the soma with 330 µM Oregon Green 488 BAPTA-1 conjugated to 70-kDa dextran (Molecular Probes, Inc.), using an Eppendorf pressure injection system (Transjector 5246). Injection solutions comprised (in mM): 5 HEPES pH 7.2, 72 KCl, 12 NaCl and 0.05 EGTA. The injection time was ~0.3 s per cell⁻¹ and a constant back pressure of 0.4–1.2 p.s.i. was applied to the micropipette. Micropipettes were pulled from thin-wall borosilicate tubing with a Flaming/Brown P-97 micropipette puller (Sutter Instr. Co.) and in some cases the tips were bevelled using a micropipette beveller (Sutter Instr. Co.). Calcium imaging was done using a BioRad confocal imaging system (MRC-1024ES), which was equipped with a 15 mW Kr/Ar ion laser and an Olympus microscope (BX50WI) with a 60× objective (LUMPlanF1; NA, 0.9). Excitation was provided by the 488-nm line of the laser attenuated to 1% and the Oregon Green BAPTA-dextran signal was detected using a 520 ± 15 nm emission filter. Images were acquired every 15 s at 512 × 512 pixels with a line scan rate of ~1 ms⁻¹. Bright-field images were collected simultaneously by a blue-sensitive photodiode. Digital images were analysed using NIH Image. The intensity of the fluorescence at the growth cone was measured over a fixed square area that covered the entire growth cone throughout the measurement period (Figs 2 and 4).

Received 30 September; accepted 16 November 1999.

1. Tessier-Lavigne, M. & Goodman, C. S. The molecular biology of axon guidance. *Science* **274**, 1123–1133 (1996).
2. Mueller, B. K. Growth cone guidance: first steps towards a deeper understanding. *Annu. Rev. Neurosci.* **22**, 351–388 (1999).
3. Song, H.-J. & Poo, M.-M. Signal transduction underlying growth cone guidance by diffusible factors. *Curr. Opin. Neurobiol.* **9**, 355–388 (1999).
4. Bixby, J. L. & Spitzer, N. C. Early differentiation of vertebrate spinal neurons in the absence of voltage-dependent Ca²⁺ and Na⁺ influx. *Dev. Biol.* **106**, 89–96 (1984).
5. Kater, S. B., Mattson, M. P., Cohan, C. & Conner, J. A. Calcium regulation of the neuronal growth cone. *Trends Neurosci.* **11**, 317–323 (1988).
6. Bandtlow, C. E., Schmidt, M. F., Hassinger, T. D., Schwab, M. E. & Kater, S. S. Role of intracellular Ca²⁺ in NI-35-evoked collapse of neuronal growth cones. *Science* **259**, 80–83 (1993).
7. Gu, X. & Spitzer, N. C. Distinct aspects of neuronal differentiation encoded by frequency of spontaneous Ca²⁺ transients. *Nature* **375**, 784–787 (1995).
8. Takei, K., Shin, R. M., Inoue, T., Kato, K. & Mikoshiba, K. Regulation of nerve growth mediated by inositol 1,4,5-trisphosphate receptors in growth cones. *Science* **282**, 1705–1708 (1998).
9. Gomez, T. M. & Spitzer, N. C. *In vivo* regulation of axon extension and pathfinding by growth-cone calcium transients. *Nature* **397**, 350–355 (1999).
10. Culotti, J. G. & Merz, D. C. DCC and netrins. *Curr. Opin. Cell Biol.* **10**, 609–613 (1998).
11. Berridge, M. J. Neuronal calcium signaling. *Neuron* **21**, 13–26 (1998).
12. Lohof, A. M., Quillan, M., Dan, Y. & Poo, M.-M. Asymmetric modulation of cytosolic cAMP activity induces growth cone turning. *J. Neurosci.* **12**, 1253–1261 (1992).
13. Ming, G. L. *et al.* cAMP-dependent growth cone guidance by netrin-1. *Neuron* **19**, 1225–1235 (1997).
14. Leis, H. J. *et al.* Inhibition of prostanoic acid formation in intact cells by 2,5-di-(tert-butyl)-1,4-benzohydroquinone, a blocker of Ca²⁺-ATPase. *Brit. J. Pharmacol.* **117**, 540–544 (1996).
15. Thastrup, O., Cullen, P. J., Drbak, B. K., Hanley, M. R. & Dawson, A. P. Thapsigargin, a tumor promoter, discharges intracellular Ca²⁺ stores by specific inhibition of the endoplasmic reticulum Ca²⁺-ATPase. *Proc. Natl Acad. Sci. USA* **87**, 2466–2470 (1990).
16. Marchetti, C., Amico, C. & Usai, C. Functional characterization of the effect of nimodipine on the calcium current in rat cerebellar granule cells. *J. Neurophysiol.* **73**, 1169–1180 (1995).
17. Zucchi, R. & Ronca-Testoni, S. The sarcoplasmic reticulum Ca²⁺ channel/ryanodine receptor: modulation by endogenous effectors, drugs and disease states. *Pharmacol. Rev.* **49**, 1–51 (1997).
18. Lin, K., Sadee, W. & Quillan, J. M. Rapid measurements of intracellular calcium using a fluorescence plate reader. *Biotechniques* **26**, 318–322 (1999).
19. Zheng, J. Q., Felder, M., Connor, J. A. & Poo, M.-M. Turning of nerve growth cones induced by neurotransmitters. *Nature* **368**, 140–144 (1994).
20. Lau, P. M., Zucker, R. S. & Bentley, D. Induction of filopodia by direct elevation of intracellular calcium ion concentration. *J. Cell Biol.* **45**, 1265–1275 (1999).
21. Zheng, J. Q. Turning of nerve growth cones induced by localized increases in intracellular calcium ions. *Nature* **403**, 89–93 (2000).
22. Song, H.-J., Ming, G. L. & Poo, M.-M. Cyclic-AMP-induced switching in turning direction of nerve growth cones. *Nature* **388**, 275–279 (1997).
23. Song, H.-J. *et al.* Conversion of neuronal growth cone responses from repulsion to attraction by cyclic nucleotides. *Science* **281**, 1515–1518 (1998).
24. Islam, M. S. *et al.* *In situ* activation of the type 2 ryanodine receptor in pancreatic beta cells requires cAMP-dependent phosphorylation. *Proc. Natl Acad. Sci. USA* **95**, 6145–6150 (1998).
25. Ming, G. *et al.* Phospholipase C-γ and phosphoinositide 3-kinase mediate cytoplasmic signalling in nerve growth cone guidance. *Neuron* **23**, 139–148 (1999).
26. Blair, L. A. C. & Marshall, J. IGF-1 modulates N and L calcium channels in a PI 3-kinase-dependent manner. *Neuron* **19**, 421–429 (1997).
27. Artola, A. & Singer, W. Long-term depression of excitatory synaptic transmission and its relationship to long-term potentiation. *Trends Neurosci.* **16**, 480–487 (1993).
28. Malenka, R. C. Synaptic plasticity in the hippocampus: LTP and LTD. *Cell* **78**, 535–538 (1994).
29. Yang, S. N., Tang, Y. G. & Zucker, R. S. Selective induction of LTP and LTD by postsynaptic [Ca²⁺]_i elevation. *J. Neurophysiol.* **81**, 781–787 (1999).
30. Spitzer, N. & Lamborghini, J. E. The development of the action potential mechanism of amphibian neurons isolated in culture. *Proc. Natl Acad. Sci. USA* **73**, 1641–1645 (1976).

Acknowledgements

This work was supported by a grant from NIH to M.-m.P. M.T.-L. is an Investigator of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to M.-m.P. (e-mail: mpoo@ucsd.edu).

Caspase-12 mediates endoplasmic-reticulum-specific apoptosis and cytotoxicity by amyloid-β

Toshiyuki Nakagawa*, Hong Zhu*, Nobuhiro Morishima†, En Li‡, Jin Xu§, Bruce A. Yankner§ & Junying Yuan*

* Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA

† Biodesign Research Group, RIKEN (the Institute of Physical and Chemical Research), 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
‡ Cardiovascular Research Center, Massachusetts General Hospital, 129 13th Street, Charlestown, Massachusetts 02119, USA

§ Department of Neurology, Harvard Medical School and Children's Hospital, Enders 260, 300 Longwood Avenue, Boston, Massachusetts 02115, USA

Apoptosis, or cellular suicide, is important for normal development and tissue homeostasis, but too much or too little apoptosis can also cause disease^{1,2}. The family of cysteine proteases, the so-called caspases, are critical mediators of programmed cell death³, and thus far 14 family members have been identified. Some of these, such as caspase-8 (refs 4, 5), mediate signal transduction downstream of death receptors located on the plasma membrane. Others, such as caspase-9 (ref. 6), mediate apoptotic signals after mitochondrial damage. Stress in the endoplasmic reticulum (ER) can also result in apoptosis⁷. Here we show that caspase-12 is localized to the ER and activated by ER stress, including disruption of ER calcium homeostasis and accumulation of excess proteins in ER, but not by membrane- or mitochondrial-targeted apoptotic signals. Mice that are deficient in caspase-12 are resistant to ER stress-induced apoptosis, but their cells undergo apoptosis in response to other death stimuli. Furthermore, we show that caspase-12-deficient cortical neurons are defective in apoptosis induced by amyloid-β protein but not by staurosporine or trophic factor deprivation. Thus, caspase-12 mediates an ER-specific apoptosis pathway and may contribute to amyloid-β neurotoxicity.

Murine caspase-12 is a member of the ICE (interleukin-1β converting enzyme) subfamily of caspases. The amino-acid sequence of caspase-12 shares high homology with murine caspase-1 (39% identity) and caspase-11 (38% identity), and with human caspase-4 (48% identity) and caspase-5 (45% identity) (ref. 8; N.M. and J.Y., unpublished data). Overexpression of caspase-12 induces apoptosis, which is inhibited by the pan-caspase inhibitor z-VAD-fmk, partially by overexpression of bcl-xL, but not by crmA (T.N., N.M. and J.Y., unpublished data). Caspase-12 is ubiquitously expressed in mouse tissues; it is expressed at high levels in muscle, liver and kidney, and at moderate levels in brain, where it is present in cortical neurons, Purkinje cells, brainstem neurons and olfactory neurons (T.N. and J.Y., unpublished data).

To determine the site of caspase-12 action, we examined the subcellular localization of caspase-12. Lysates obtained from brain cortices of newborn mice were fractionated into nuclear, mitochondrial, microsomal and soluble fractions and analysed by western blotting. The quality of fractionation was controlled by the presence

of known compartment-specific proteins: cytochrome *c* for mitochondria and TRAP α for the ER. Procaspase-12 (relative molecular mass 60,000 (M_r , 60K)) was detected predominantly in the microsomal fraction (Fig. 1A), suggesting that caspase-12 may be associated with ER, whereas cleaved caspase-12, which may be produced during the fractionation procedure, was present in the soluble fraction. In contrast, procaspase-3 was detected in mitochondrial and soluble fractions, but not in the microsomal fraction.

To determine whether caspase-12 is indeed associated with ER, we examined the subcellular distribution of caspase-12, as compared with other caspases, by immunocytochemistry. Antibodies

specific to caspase-1 (Fig. 1Bd) or caspase-3 (Fig. 1Bc) diffusely stained the cytoplasm of L929 cells, whereas antibody specific to caspase-8 (Fig. 1Bb) stained both the cytoplasm and plasma membrane. In contrast, a monoclonal antibody specific to caspase-12 stained the perinuclear region of L929 cells (Fig. 1Ba) and HeLa cells (Fig. 1Bg), which is a pattern similar to that of grp78 (Bip), a chaperone protein localized to the ER⁹ (Fig. 1Be). The localization of caspase-12 in the ER was confirmed by co-staining with anti-Bip (Fig. 1Bf), as well as anti-presenilin-2 (Fig. 1Bh, k) and anti- β -amyloid precursor protein (Fig. 1Bi, l). A significant fraction of presenilin-2 (ref. 10) and β -amyloid precursor protein¹¹

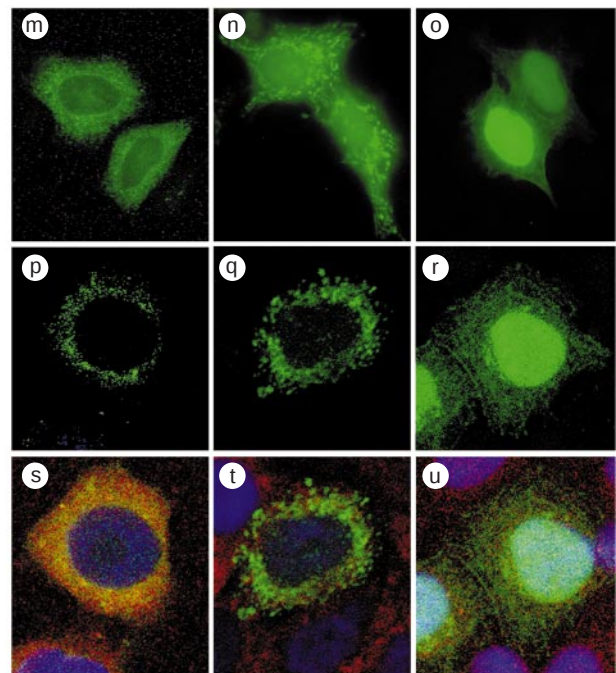
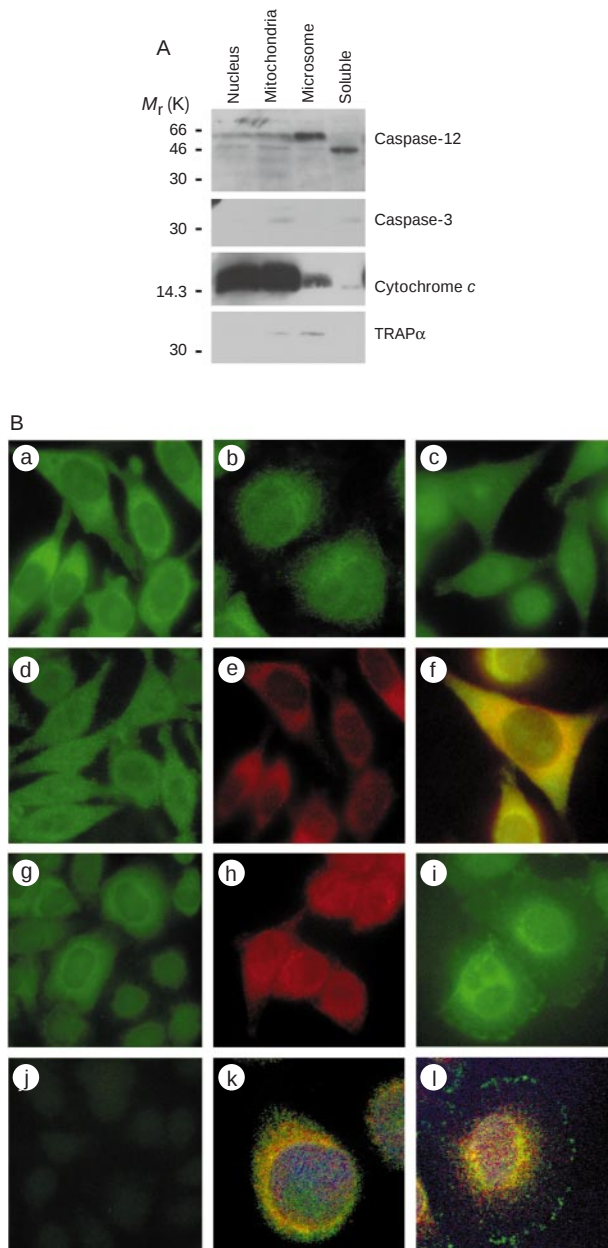


Figure 1 Caspase-12 predominantly localizes in the ER. **A**, Western blots of fractionated mouse cerebral cortical lysate immunostained by anti-caspase-12, anti-caspase-3, anti-cytochrome *c* and anti-TRAP α . M_r , relative molecular mass. **B**, Immunocytochemistry of endogenous caspase-12 (**a**), caspase-3 (**c**), caspase-1 (**d**), Bip (**e**) and merger of caspase-12 (green) and Bip (red) (**f**) in L929 cells; caspase-8 (**b**), caspase-12 (**g**), presenilin-2 (**h**) and β -amyloid precursor protein (**i**) in HeLa cells; and colocalization of caspase-12 (green) and presenilin-2 (red) (**k**), or caspase-12 (red) and β -amyloid precursor protein (green) (**l**) by confocal microscopy (nuclei were stained blue by bisbenzamide) in HeLa cells. **j**, Negative control using supernatant of a hybridoma, which

does not produce specific antibody, as primary antibody. **m–o**, Fluorescent micrographs of HeLa cells transfected with pEYFP-ER (**m**), which specifically labels ER; pEYFP-Mito (**n**), which specifically labels mitochondria; or pEYFP (**o**), which labels both cytosol and nucleus. **p–r**, Confocal images of enhanced yellow fluorescence of HeLa cells transfected with pEYFP-ER (**p**), pEYFP-Mito (**q**) and pEYFP (**r**). **s–u**, Colocalization of caspase-12 with yellow fluorescence of pEYFP-ER, but not that of pEYFP-Mito or pEYFP. HeLa cells transfected with pEYFP-ER (**s**), pEYFP-Mito (**t**) or pEYFP (**u**) were immunostained by anti-caspase-12 antibody (red).

are associated with the ER. To further confirm the ER localization of caspase-12, we transfected HeLa cells with pEYFP-ER (Fig. 1Bm, p), pEYFP-Mito (Fig. 1Bn, q) and pEYFP (Fig. 1Bo, r), which label ER, mitochondria and cytoplasm/nucleus, respectively. Co-immunostaining with anti-caspase-12 showed clearly that caspase-12 colocalized with pEYFP-ER but not with pEYFP-Mito or pEYFP

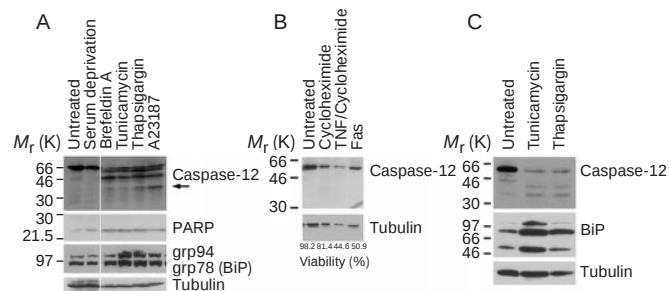


Figure 2 Caspase-12 is specifically activated by ER stress. **A**, PC12 cells were deprived of serum, or treated with brefeldin A ($10 \mu\text{g ml}^{-1}$), tunicamycin ($1 \mu\text{g ml}^{-1}$), thapsigargin ($2 \mu\text{M}$) or A23187 ($2 \mu\text{M}$) for 36 hours. Caspase-12 is specifically cleaved in cells that are undergoing ER stress-induced apoptosis. ER stress response is indicated by the induction of grp 94 and grp78 (BiP). Activation of caspase(s) is indicated by the detection of 23K fragment of poly(ADP-ribose) polymerase (PARP). **B**, Caspase-12 is not cleaved in W4 cells that have been induced to undergo apoptosis by treatment of cycloheximide ($1 \mu\text{g ml}^{-1}$), TNF (100 ng ml^{-1}) and cycloheximide ($1 \mu\text{g ml}^{-1}$) or anti-Fas ($4 \mu\text{g ml}^{-1}$) for 8 hours. **C**, Caspase-12 is cleaved in embryonic fibroblast cells that have been induced to undergo apoptosis by treatment of tunicamycin ($1 \mu\text{g ml}^{-1}$) and thapsigargin ($2 \mu\text{M}$).

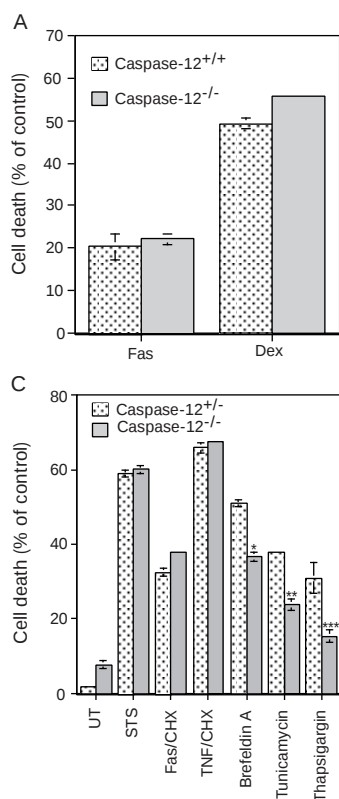
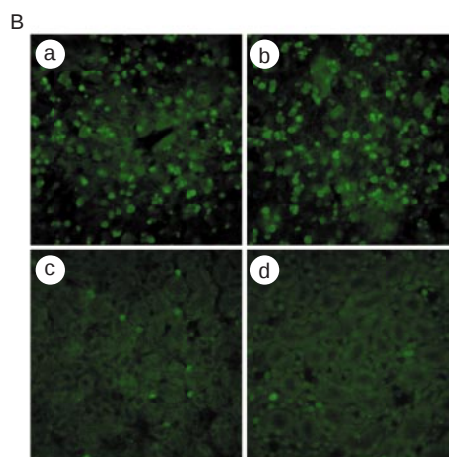


Figure 3 Caspase-12 is involved in ER stress-induced apoptosis, but not in non-ER stress-induced apoptosis. **A**, Thymocytes from wild-type (caspase-12^{+/+}) and mutant (caspase-12^{-/-}) mice die equally after treatment of anti-Fas (Jo2, $1 \mu\text{g ml}^{-1}$) or dexamethasone (Dex, $2 \mu\text{M}$) for 24 hours. Cell viability was determined by MTT assay. **B**, Apoptosis induced by anti-Fas antibody *in vivo* is not different in caspase-12^{+/+} and caspase-12^{-/-} mice. Mice were injected with anti-Fas antibody (Jo2, $100 \mu\text{g}$ per mouse) intraperitoneally and the percentages of apoptotic cells in liver (**a**, **b**) and kidney (**c**, **d**) were analysed by TUNEL staining. A quantitative analysis showed that the number of TUNEL positive cells per high-power field was 92 ± 23 in caspase-12^{+/+} ($n = 3$ mice)

(Fig. 1Bs–u). Together, these results indicate that caspase-12 is predominantly associated with ER.

To determine the specificity of the caspase-12 pathway, we examined the activation of caspase-12, as indicated by cleavage of procaspase-12, in response to different apoptotic stimuli. Procaspase-12 was not activated in PC12 cells that were induced to undergo apoptosis by serum deprivation (Fig. 2A), or in W4 cells that have been induced to die by CHX, tumour necrosis factor (TNF) and CHX or anti-Fas alone (Fig. 2B). In contrast, treatment of PC12 cells with brefeldin A, tunicamycin, thapsigargin and calcium ionophore A23187, or treatment of embryonic fibroblast cells with tunicamycin and thapsigargin induced effective cleavage of procaspase-12. Brefeldin A, an inhibitor of ER–Golgi transport, and tunicamycin, a specific inhibitor of *N*-glycosylation in ER, are effective inducers of ER stress as indicated by the upregulation of grp94 and grp78 (BiP)⁹ (Fig. 2A). The calcium ionophore A23187 and thapsigargin disrupted intracellular calcium homeostasis and also induced ER stress (Fig. 2A, C). As cleavage of caspases is a common indicator of activation, we conclude that caspase-12 can be specifically activated by apoptotic signals with an ER stress component, but not by apoptotic signals that do not induce ER stress.

To address the function of caspase-12 *in vivo*, we generated caspase-12-null mutant mice in which a *neo* gene has replaced a segment of the intron III and the exon IV of caspase-12, making the downstream sequence out of the reading frame. Caspase-12 gene transcription and translation products could not be detected in mutant mice by western blotting or polymerase chain reaction with reverse transcriptase (RT-PCR) (data not shown). Mice homozy-



and 115 ± 12 in caspase-12^{-/-} liver ($n = 3$ mice) (two-tailed P value = 0.41); and 6 ± 3 in caspase-12^{+/+} ($n = 3$ mice) and 6 ± 1 in caspase-12^{-/-} kidney ($n = 3$ mice) (two-tailed value = 0.48). **C**, Caspase-12^{-/-} embryonic fibroblast cells are resistant to ER stress-induced apoptosis, not to Fas-TNF- or STS-induced cell death. Embryonic fibroblast cells were treated with STS ($2 \mu\text{M}$), anti-Fas (Jo2, $4 \mu\text{g ml}^{-1}$) plus cycloheximide (CHX, $10 \mu\text{g ml}^{-1}$) or TNF (100 ng ml^{-1}) plus cycloheximide (CHX, $10 \mu\text{g ml}^{-1}$) for 8 hours, or brefeldin A ($10 \mu\text{g ml}^{-1}$), tunicamycin ($1 \mu\text{g ml}^{-1}$) or thapsigargin ($2 \mu\text{M}$) for 24 hours. Cell viability was assayed by LIVE/DEAD viability/cytotoxicity kit (Molecular Probes). Two-tailed P value: asterisk, $P = 0.0001$; two asterisks, $P = 0.0002$; three asterisks, $P = 0.0022$.

gous for the germline caspase-12 mutation (caspase-12^{-/-}) were indistinguishable in appearance from wild-type littermates. Histological analysis of caspase-12 did not detect any developmental abnormalities. The ratio of caspase-12^{-/-} to caspase-12^{+/-}/caspase-12^{+/+} from a heterozygous cross was consistent with the predicted mendelian ratio (data not shown).

To determine the role of caspase-12 in apoptosis, we first examined apoptosis of thymocytes isolated from wild-type and caspase-12^{-/-} mice induced by anti-Fas or dexamethasone. Caspase-12^{+/+} and caspase-12^{-/-} thymocytes showed similar levels of cell death after treatment with anti-Fas or dexamethasone for 24 hours (Fig. 3A), indicating that caspase-12 is not essential for apoptosis induced by anti-Fas or dexamethasone in thymocytes. We injected

anti-Fas antibody intraperitoneally in mice to examine whether caspase-12 is needed for the Fas apoptosis pathway *in vivo*. Three hours later, liver and kidney were collected and apoptosis of liver and kidney cells were examined by terminal deoxynucleotidyl transferase (TdT)-mediated dUTP nick end labelling (TUNEL) staining. The percentages of TUNEL-positive cells were not different in liver and kidney of wild-type and caspase-12^{-/-} mice (Fig. 3B). Thus, caspase-12 is not essential for anti-Fas-induced apoptosis *in vivo*. As we found that caspase-12 is specifically activated by ER stress signals in embryonic fibroblast cells, we examined whether caspase-12 is required specifically for ER stress-induced apoptosis. Embryonic fibroblast cells were treated either with non-ER stress-inducing apoptotic stimuli, such as staurosporine, anti-Fas plus cycloheximide and TNF plus cycloheximide, or with ER stress-inducing apoptotic stimuli, such as brefeldin A, tunicamycin and thapsigargin. Caspase-12^{-/-} embryonic fibroblast cells were more resistant to ER stress-inducing apoptotic signals than caspase-12^{+/+} embryonic fibroblast cells, but caspase-12^{-/-} and caspase-12^{+/+} embryonic fibroblast cells showed similar sensitive to non-ER stress-induced apoptotic signals (Fig. 3C). These results indicate that caspase-12 contributes to the mechanism of ER stress-induced apoptosis, but that it is not involved in non-ER stress-induced apoptosis.

Tunicamycin has been reported to induce ER stress-mediated apoptosis in renal epithelial cells¹². Because caspase-12 is specifically expressed in renal proximal tubular epithelial cells, but not in the glomerulus (Fig. 4A), we examined the involvement of caspase-12 in renal tubular epithelial cell death after tunicamycin treatment. Sublethal doses of tunicamycin (0.25–1 µg g⁻¹ weight) were injected intraperitoneally into wild-type and caspase-12^{-/-} mice. Although both wild-type and caspase-12^{-/-} mice suffered significant weight loss in response to tunicamycin injection (data not shown) at the time of experimental termination, two out of five wild-type mice, but none of the caspase-12^{-/-} mice, died 4 days after injection (1 µg g⁻¹). Histological examination of the kidneys from wild-type mice showed extensive haemorrhagic lesions (Fig. 4Ba–d). Such histological alterations were restricted to the tubular epithelium, whereas glomeruli were mostly spared. Renal tubular epithelium of wild-type mice was completely destroyed after injection of high-dose tunicamycin (1 µg g⁻¹); in contrast, normal renal tubular epithelial cells of caspase-12^{-/-} mice were mostly preserved when injected with the same dose (Fig. 4Be). TUNEL staining of wild-type kidney 4 days after tunicamycin (0.25 µg g⁻¹) treatment showed a large number of TUNEL-positive cells, whereas significantly less TUNEL-positive cells were found in the caspase-12^{-/-} kidney (Fig. 4Bf, g). Quantitative analysis showed that the number of TUNEL-positive cells per high-power field was 149 ± 41 in caspase-12^{+/+} or caspase-12^{+/-} (n = 4) and 15 ± 5.8 in caspase-12^{-/-} kidney (n = 4) (two-tailed P value < 0.03). These results indicate that caspase-12 is critical in mediating tunicamycin-induced renal tubular epithelial cell death.

The amyloid-β (Aβ) protein has been shown to induce neuronal cell death, which may contribute to the mechanism of neuronal degeneration in Alzheimer's disease¹³. As the ER is a potential intracellular target of Aβ protein, and a putative intracellular Aβ receptor, ERAB, is an ER membrane protein¹⁴, we determined whether caspase-12 could contribute to the mechanism of Aβ toxicity. To determine the involvement of caspase-12 in Aβ neurotoxicity, we assayed Aβ neurotoxicity in primary cortical neurons isolated from wild-type, caspase-12^{-/-} and caspase-11^{-/-} (ref. 15; used as a control) mice using the percentages of neurons with condensed nuclei as an indicator of apoptosis (Fig. 5A, B). 24.9 (± 1.6) % of wild-type compared with 8.6 (± 1.6) % of caspase-12^{-/-} neurons showed apoptotic morphology after incubation with 20 µM fibrillar Aβ protein (1–40) for 72 hours (P < 0.0002). In contrast, the percentage of apoptotic caspase-11^{-/-} neurons (22.8%) was not significantly different from that of wild type. To determine

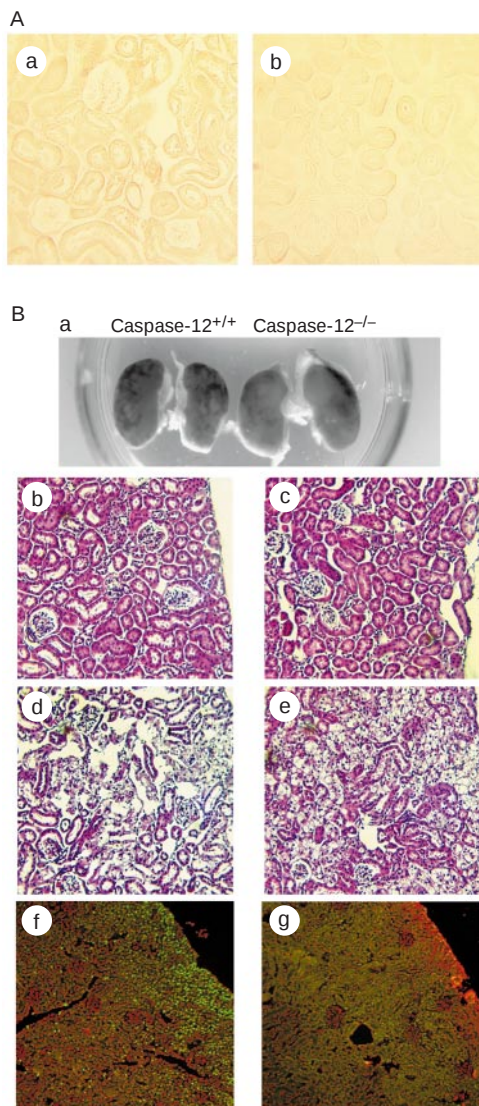


Figure 4 Inhibition of tunicamycin induced renal tubular damage by loss of caspase-12 function. **A**, Immunohistochemistry of caspase-12 in kidneys of caspase-12^{+/+} (**a**) and caspase-12^{-/-} mice (**b**). Caspase-12 is mainly expressed in renal tubular cells. **B**, Inhibition of tunicamycin-induced apoptosis and cellular damage by a mutation in caspase-12. **a**, Four per cent paraformaldehyde-perfused kidneys of wild-type and caspase-12^{-/-} mice. **b–e**, Wild-type (**b**, **d**) and caspase-12^{-/-} (**c**, **e**) kidneys after tunicamycin injection. Haematoxylin and eosin stained kidney samples shown are untreated (**b**, **c**), and 4 days after injection of 1 µg g⁻¹ tunicamycin (**d**, **e**). **f**, **g**, TUNEL staining of kidney samples 4 days after injection of 0.25 µg g⁻¹ tunicamycin. Caspase-12^{+/+} (**f**); caspase-12^{-/-} (**g**). The number of TUNEL-positive cells per high-powered field was counted in four caspase-12^{+/+} or caspase-12^{+/-} and four caspase-12^{-/-} tunicamycin-treated mice.

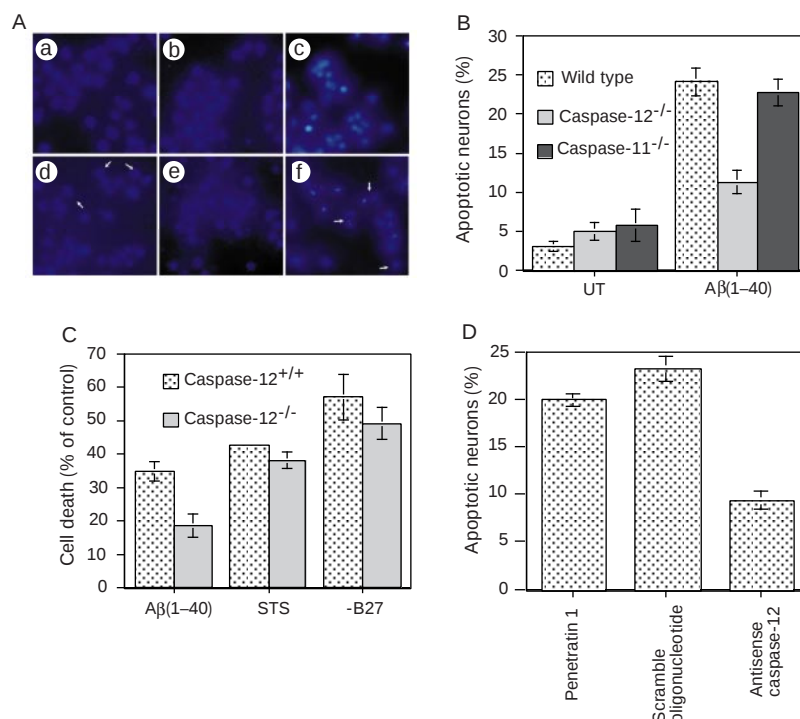


Figure 5 Caspase-12 is required for A β neurotoxicity in cortical neurons. **A**, Hoechst 33258 staining of cortical neurons that were either untreated (**a, b, c**) or exposed to 20 μ M A β protein (1–40) for 72 hours (**d, e, f**). Arrows indicate typical apoptotic nuclei in caspase-12^{+/+} and caspase-11^{-/-} neurons. Wild-type (**a, d**), caspase-12^{-/-} (**b, e**) and caspase-11^{-/-} (**c, f**) neurons are shown. **B**, Quantitative analysis of neuronal apoptosis induced by A β protein (1–40). Apoptotic neurons were scored as condensed nuclei after Hoechst 33258 staining. 24.9 ($\pm$ 1.6) % of wild-type, 8.6 ($\pm$ 1.6) % of caspase-12^{-/-} and 22.8 ($\pm$ 1.7) % of caspase-11^{-/-} neurons died after incubation with 20 μ M fibrillar A β protein (1–40) for 72 hours. P < 0.0002 (wild-type versus caspase-12^{-/-}) and P < 0.0001 (caspase-12^{-/-} versus caspase-11^{-/-}). UT, untreated. **C**, Neurons were treated with A β protein (20 μ M A β (1–40)) or staurosporine (1.0 μ M STS) or were

deprived B27 supplement (–B27; Gibco BRL). Cell viability was determined by MTS assay after 72 hours of A β , 48 hours of STS treatment, or 4 days of B27 deprivation. 35.0 ($\pm$ 2.9) % of wild-type neurons compared with 19.0 ($\pm$ 3.5) % of caspase-12^{-/-} neurons died after incubating with A β protein (1–40) for 72 hours (P < 0.01). Values are mean $\pm$ s.e.m. The same statistical significance was obtained in three independent experiments. **D**, Antisense oligonucleotides of caspase-12 protect against A β neurotoxicity. Wild-type cortical neurons were treated with penetratin 1, penetratin 1 plus antisense of caspase-12, or penetratin 1 plus scramble oligonucleotides for 6 hours, and exposed to A β protein (20 μ M A β (1–40)) for 3 days. Antisense oligonucleotides were added every 24 hours. Apoptotic neurons were scored as condensed nuclei with Hoechst 33258 staining.

whether loss of caspase-12 offers a general protection against all apoptotic stimuli, we induced apoptosis of cortical neurons by treatment with A β protein, staurosporine or trophic factor deprivation. Although caspase-12^{-/-} cortical neurons are more resistant to A β protein neurotoxicity than wild-type neurons, they are not protected against apoptosis induced by staurosporine or trophic factor deprivation (Fig. 5C). To exclude the possibility that the apparent protection against A β protein neurotoxicity was due to the loss of a population of A β -protein-sensitive neurons during the development of caspase-12^{-/-} mice, we determined whether treatment with antisense caspase-12 can protect wild-type cortical neurons. Wild-type cortical neurons were treated for 6 hours with antisense caspase-12, or with a scrambled control oligonucleotide in the presence of penetratin 1 to facilitate the entry of oligos¹⁶, and then exposed to A β protein for 3 days. Cortical neurons that have been treated with antisense caspase-12 showed significantly less cell death than those that had been treated with either penetratin 1 alone, or penetratin 1 plus the scrambled control oligo. These results indicate that caspase-12 is specifically required for apoptosis induced by A β protein and other ER stress-inducing signals, but is not generally involved in other apoptotic pathways.

Thus, we have identified a new caspase-12-mediated apoptosis pathway in the ER. Although other caspases may be activated downstream from caspase-12, caspase-12 appears to be necessary for apoptosis induced by a variety of ER-directed pro-apoptosis signals, including A β neurotoxicity in cortical neurons. As caspase-12^{-/-} mice do not exhibit any obvious developmental or behavioural defects and do not show an increased incidence of cancer, our results

suggest that the human orthologue of caspase-12 may be a potential therapeutic target for Alzheimer's disease. □

Methods

Materials

We used the following antibodies and chemicals. Cytochrome *c* antibody and anti-mouse Fas monoclonal antibody (Jo2) (Pharmingen); BiP antibody (StressGen); FITC or Texas Red anti-rat, TRITC anti-mouse or anti-Rabbit, FITC anti-Rabbit IgG (H+L) antibody (Jackson ImmunoResearch Laboratories); tubulin antibody (Sigma); PARP antibody (R&D systems); presenilin-2 antibody (Santa Cruz); and β -amyloid precursor protein antibody (ZYMED); zVAD.fmk (Bachem); brefeldin A, tunicamycin, A23187, cycloheximide and staurosporine (Sigma); thapsigargin (Research Biochemical International); TNF- α (R&D systems); and penetratin 1 (Oncor). Anti-caspase-12 was made by injecting rat with a His-tagged fusion protein of caspase-12 (residues 95–318) expressed in pET15b.

Fractionation experiments

Fractionation was done as described¹⁷. Briefly, brain cortices of newborn mice were lysed and homogenized in buffer A (50 mM Tris-HCl pH 8.0, 1 mM β -mercaptoethanol, 1 mM EDTA, 0.32 M sucrose and 0.1 mM PMSF). Nuclear fraction was the first, 900g, 10-min pellet, and mitochondrial fraction was the second, 5000g, 10-min pellet lysed by buffer A with equal volume of 2 $\times$ SDS buffer (125 mM Tris-HCl pH 6.8, 4% SDS, 1.44 M β -mercaptoethanol, 14% glycerol and 0.2% bromophenol blue). Microsomal and soluble fractions were extracted from the 105,000g, 60-min pellet, and supernatant, respectively. Western blotting was carried out using the manufacturers' protocols (Immobilon-P; Millipore, ECL; Amersham) using anti-caspase-12, anti-caspase-3 (a gift from J. L. Goldstein), anti-cytochrome *c* and anti-TRAP α (a gift from T. Rapoport).

Cell culture, transfection and immunocytochemistry

Endogenous expression of caspases-12, -3, -1 and -8, BiP, presenilin 2 and amyloid precursor protein in L929 cells or HeLa cells was detected by immunocytochemistry. Cells were fixed in 4% paraformaldehyde for 15 min at 4°C. Immunostaining was done by using anti-caspase-12, -8, -3, -1, anti-BiP, anti-presenilin-2 and anti- β -amyloid precursor

protein antibodies, which were followed by Fluorescein (FITC)-, Rhodamine (TRITC)-, or Texas Red-conjugated anti-rat, anti-rabbit or anti-mouse antibodies, and nuclear staining (bisbenzimidazole). Samples were examined with a fluorescence or AXIOVERT 135 TV microscope (Zeiss). HeLa cells were seeded on cover glasses coated by poly-L-lysine and transfected with 2.0 μ g expression plasmid DNA (pEYFP-ER, pEYFP-Mito, pEYFP; CLONTECH) using calcium phosphate precipitation. Cells were fixed in 4% paraformaldehyde, immunostained with monoclonal caspase-12 antibody and examined with a confocal microscope (Zeiss).

Generation of caspase-12 knockout mice

Gene targeting was done essentially as described¹⁷. Briefly, the caspase-12 gene fragments were cloned from a 129/sv mouse genomic library (Stratagene). DT-A fragment (pMCIDT-A; a gift from A. Yagi) was cloned into pGEM7 vector with a neomycin-resistant gene (PGK-neo). A 2.0-kilobase (kb) *SacI*-*StuI* fragment and a 9.0-kb *NofI* (cloning site from phage vector)-*XbaI* fragment were subcloned into PGK-neo-DT-A vector to generate the targeting construct. The targeting construct was transfected into J-1 ES cells by electroporation. G418 resistant transfectants were screened by genomic Southern blot analysis with a 3' flanking probe. The efficiency of homologous recombination was 5 in 232 clones. Two of the targeting clones transmitted the mutant allele. Homozygous mice were generated by interbreeding heterozygous mice (129 $\times$ C57BL/6J).

Histology and TUNEL assay of kidney and liver

Caspase-12^{+/+}, caspase-12^{+/-} and caspase-12^{-/-} littermates were injected with tunicamycin (0.25–1.0 μ g g⁻¹ weight) intraperitoneally and killed 4 days after injection. Kidneys were fixed in 4% paraformaldehyde and embedded in paraffin. Histological analysis of kidney was done by haematoxylin and eosin staining. Four-week-old caspase-12^{+/+} and caspase-12^{-/-} mice were injected with Fas antibody (Jo2, 100 μ g per mouse) intraperitoneally and killed 3 hours after injection. Liver and kidney were fixed in 4% paraformaldehyde and embedded in paraffin. TUNEL assay was done by manufacturers' protocols (ApopTag; Oncor).

Thymocytes culture

Thymus was dissected and passed through mesh (200–300 μ m). Cells were collected by centrifugation (160g, 5 min) and resuspended in culture medium containing RPMI and 10% fetal bovine serum. Thymocytes were treated with Fas antibody (Jo2, 1 μ g ml⁻¹) or dexamethasone (2 μ M) for 12 hours. Percentage of cell-death was analysed by MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide; thiazolyl blue).

Embryonic fibroblast cells

Mouse embryonic fibroblast cells were prepared as described¹⁷. Briefly, homozygous caspase-12^{-/-} females mated with heterozygous males were killed at day 14.5 of gestation. Embryos were decapitated and eviscerated and carcasses were digested in trypsin. Cells were split and frozen at passage 2.

Induction of cell death

Embryonic fibroblast cells (8 $\times$ 10³ cells ml⁻¹) were treated with staurosporine (2 μ M), anti-Fas (Jo2, 4 μ g ml⁻¹) plus CHX (10 μ g ml⁻¹), or TNF (100 ng ml⁻¹) plus CHX (10 μ g ml⁻¹) for 8 hours, Brefeldin A (10 μ g ml⁻¹), tunicamycin (1 μ g ml⁻¹) or thapsigargin (2 μ M) for 24 hours. Cell death was assayed by the LIVE/DEAD viability/cytotoxicity kit.

Preparation of cell lysates and western blotting

PC12 cells treated by several drugs or serum deprivation for 36 hours were collected by scraping and lysed in lysis buffer (62.5 mM Tris-HCl pH 6.8, 2% SDS, 0.72 M β -mercaptoethanol and 7% glycerol). Lysates were analysed by western blotting using 1:50 dilution of caspase-12 antibody, 1:2500 dilution of PARP antibody, 1:10,000 dilution of tubulin antibody.

Primary cortical neuron culture and A β treatment

Cerebral cortex of mouse embryo at day 14.5–15.5 was freed from meninges and separated from olfactory bulb and hippocampus. Trypsinized cells were suspended in medium (DMEM/F-12, B27 (Gibco), and penicillin, streptomycin) and seeded on poly-L-lysine-coated dishes. After 4 days of culturing, neurons were incubated with either 20 μ M A β (1–40) for 72 hours or 1.0 μ M STS for 48 hours. Cell viability was determined by MTS (3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulphenyl)-2H-tetrazolium, inner salt) assay (CellTiter 96R aqueous non-radioactive cell proliferative assay; Promega). Alternatively, apoptotic neurons were identified by characteristic nuclear morphology after fixing in 4% paraformaldehyde for 15 min at 4°C, permeabilizing in 0.4% Triton for 10 min at room temperature and staining in Hoechst 33258 (1 μ g ml⁻¹).

Antisense caspase-12 treatment

Coupling and treatment of penetratin 1 and antisense of caspase-12 were prepared as described¹⁶. Antisense caspase-12: 5'-thiol modification-TGTCCTCTGCGCCGCATG GCTGT-3'-amino linker; scramble oligonucleotide: 5'-thiol modification-GTCGCTCTGTGACGCCTGTGCTG-3'-amino linker.

Received 19 October; accepted 19 November 1999.

- Los, M., Wesselborg, S. & Schulze-Osthoff, K. The role of caspases in development, immunity, and apoptotic signal transduction: lessons from knockout mice. *Immunity* **10**, 629–639 (1999).

- Friedland, R. M. & Yuan, J. ICE, neuronal apoptosis and neurodegeneration. *Cell Death Differ.* **5**, 823–831 (1998).
- Cryns, V. & Yuan, J. Proteases to die for. *Genes Dev.* **12**, 1551–1570 (1998).
- Muzio, M. *et al.* FLICE, a novel FADD-homologous ICE/CED-3-like protease, is recruited to the CD95 (Fas/APO-1) death inducing signaling complex. *Cell* **85**, 817–827 (1996).
- Boldin, M. P., Goncharov, T. M., Goltsev, Y. V. & Wallach, D. Involvement of MACH, a novel MORT1/FADD-interacting protease, in Fas/APO-1- and TNF receptor-induced cell death. *Cell* **85**, 803–815 (1996).
- Li, P. *et al.* Cytochrome c and dATP-dependent formation of Apaf-1/caspase-9 complex initiates an apoptotic protease cascade. *Cell* **91**, 479–489 (1997).
- Welihinda, A. A., Tirasophon, W. & Kaufman, R. J. The cellular response to protein misfolding in the endoplasmic reticulum. *Gene Expr.* **7**, 293–300 (1999).
- Van de Craen, M. *et al.* Characterization of seven murine caspase family members. *FEBS Lett* **403**, 61–69 (1997).
- Kozutsumi, Y., Segal, M., Normington, K., Gething, M. J. & Sambrook, J. The presence of misfolded proteins in the endoplasmic reticulum signals the induction of glucose-regulated proteins. *Nature* **332**, 462–464 (1988).
- Walter, J. *et al.* The Alzheimer's disease-associated presenilins are differentially phosphorylated proteins located predominantly within the endoplasmic reticulum. *Mol. Med.* **2**, 673–691 (1996).
- Gabuzda, D., Busciglio, J., Chen, L. B., Matsudaira, P. & Yankner, B. A. Inhibition of energy metabolism alters the processing of amyloid precursor protein and induces a potentially amyloidogenic derivative. *J. Biol. Chem.* **269**, 13623–13628 (1994).
- Zinsner, H. *et al.* CHOP is implicated in programmed cell death in response to impaired function of the endoplasmic reticulum. *Gene Dev.* **12**, 982–995 (1998).
- Yankner, B. A., Duffy, L. K. & Kirschner, D. A. Neurotrophic and neurotoxic effects of amyloid beta protein: reversal by tachykinin neuropeptides. *Science* **250**, 279–282 (1990).
- Yan, S. D. *et al.* An intracellular protein that binds amyloid-beta peptide and mediates neurotoxicity in Alzheimer's disease. *Nature* **389**, 689–695 (1997).
- Wang, S. *et al.* Murine caspase-11, an ICE interacting protease, is essential for the activation of ICE. *Cell* **92**, 501–509 (1998).
- Troy, C. M., Derosi, D., Prochiantz, A., Greene, L. A. & Shelanski, M. L. Downregulation of Cu/Zn superoxide dismutase leads to cell death via the nitric oxide-peroxynitrite pathway. *J. Neurosci.* **16**, 253–261 (1996).
- Matsumoto, M. *et al.* Ataxia and epileptic seizures in mice lacking type I inositol 1,4,5-trisphosphate receptor. *Nature* **379**, 168–171 (1996).

Acknowledgements

We thank Y.-K. Jung and B. Hyman for helpful suggestions; A. Yagi for pMCIDT-A; T. Noda for PGK-neo plasmids; J. L. Goldstein for anti-caspase-3; T. Rapoport for anti-TRAP α ; S. Nagata for W4 cells; M. Yuan for technical help in the preparation of A β , and S.-J. Kang for providing caspase-11 mutant mice. This work was supported in part by grants from Hoechst-Marion-Roussel and Harvard Medical School project (to J.Y.) and from Human Frontier Science Program Organization and TOYOBO Biotechnology Foundation (to T.N.).

Correspondence and requests for materials should be addressed to J.Y. (e-mail: jyuan@hms.harvard.edu).

Anti-inflammatory cyclopentenone prostaglandins are direct inhibitors of I κ B kinase

Antonio Rossi^{*†}, Pankaj Kapahi^{†‡}, Gioacchino Natoli^{†‡}, Takayuki Takahashi[‡], Yi Chen[‡], Michael Karin^{†‡} & M. Gabriella Santoro^{*†§}

^{*} Institute of Experimental Medicine, Italian National Council of Research, and [§] Laboratory of Virology, Department of Biology, University of Rome Tor Vergata, 00133 Rome, Italy

[‡] Laboratory of Gene Regulation and Signal Transduction, Department of Pharmacology, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093-0636, USA

[†] These authors contributed equally to this work

NF- κ B is a critical activator of genes involved in inflammation and immunity^{1,2}. Pro-inflammatory cytokines activate the I κ B kinase (IKK) complex that phosphorylates the NF- κ B inhibitors, triggering their conjugation with ubiquitin and subsequent degradation^{3,4}. Freed NF- κ B dimers translocate to the nucleus and induce target genes, including the one for cyclo-oxygenase 2 (COX2), which catalyses the synthesis of pro-inflammatory prostaglandins, in particular PGE^{5,6}. At late stages of inflammatory episodes, how-

ever, COX2 directs the synthesis of anti-inflammatory cyclopentenone prostaglandins, suggesting a role for these molecules in the resolution of inflammation⁷. Cyclopentenone prostaglandins have been suggested to exert anti-inflammatory activity through the activation of peroxisome proliferator-activated receptor- γ (refs 8, 9). Here we demonstrate a novel mechanism of anti-inflammatory activity which is based on the direct inhibition and modification of the IKK β subunit of IKK. As IKK β is responsible for the activation of NF- κ B by pro-inflammatory stimuli^{10,11}, our findings explain how cyclopentenone prostaglandins function and can be used to improve the utility of COX2 inhibitors.

Cyclopentenone prostaglandins (cyPGs) have anti-inflammatory^{12,13} and antiviral activity^{12,13}, and inhibit NF- κ B activation in human cells stimulated with tumour necrosis factor- α (TNF α) or 12-*O*-tetradecanoylphorbol-13-acetate (TPA) by preventing phosphorylation and degradation of the NF- κ B inhibitor I κ B α ¹⁴. I κ B α is phosphorylated by the IKK complex, which contains two catalytic subunits (IKK α and IKK β) and the IKK γ or NEMO regulatory subunit^{15,16}, at sites that trigger its ubiquitin-dependent degradation^{3,4}.

To determine whether IKK could be the target for cyPGs, we treated human lymphoblastoid Jurkat cells with the cyPG prostaglandin A₁ (PGA₁) and stimulated them with either TNF α (Fig. 1a) or TPA (Fig. 1b). Both TNF α and TPA rapidly stimulated IKK activity, which reached a maximum after 10–15 min. PGA₁ inhibited IKK activity and prevented I κ B α degradation and NF- κ B

activation by both inducers (Fig. 1a, b), but did not inhibit activation of *c*-Jun amino-terminal kinases (JNK) and p38 mitogen-activated protein (MAP) kinases (Fig. 1c). Hence, PGA₁ targets a selected subset of kinases that activate transcription factors. In addition, PGA₁ did not inhibit the DNA binding of transcription factors Oct-1 and Sp-1 (Fig. 1a). A 15-min exposure to PGA₁ was sufficient to partially inhibit IKK activity, but more substantial inhibition required longer pre-incubation (Fig. 1d). Similarly, PGA₁ inhibited IKK activity and prevented NF- κ B activation in HeLa cells stimulated with interleukin1 (IL-1) or TNF α (Fig. 1e). Thus, the inhibitory effect of PGA₁ is independent of type of cell and stimulus. However, the half-maximal inhibitory (IC₅₀) concentration of PGA₁ towards NF- κ B was 12 μ M in Jurkat cells, whereas the IC₅₀ was 65 μ M in HeLa cells (data not shown).

IKK can be activated by the overexpression of protein kinases MEKK1 (MAP kinase Erk kinase 1) or NIK (NF- κ B-inducing kinase)¹⁷. When added 10 min or 18 h after the transfection of COS cells with X-press-tagged NIK, PGA₁ inhibited activation of endogenous IKK activity (Fig. 2a). In a second experiment, COS cells were co-transfected with haemagglutinin A(HA)-tagged IKK α together with either 'empty' or NIK vectors. In the conditions that we used, HA-IKK α is incorporated into functional cytokine-responsive complexes with relative molecular mass (*M*_r) 900K¹⁵. NIK-induced IKK activity was inhibited by PGA₁ added immediately after or 18 h after transfection (Fig. 2b). PGA₁

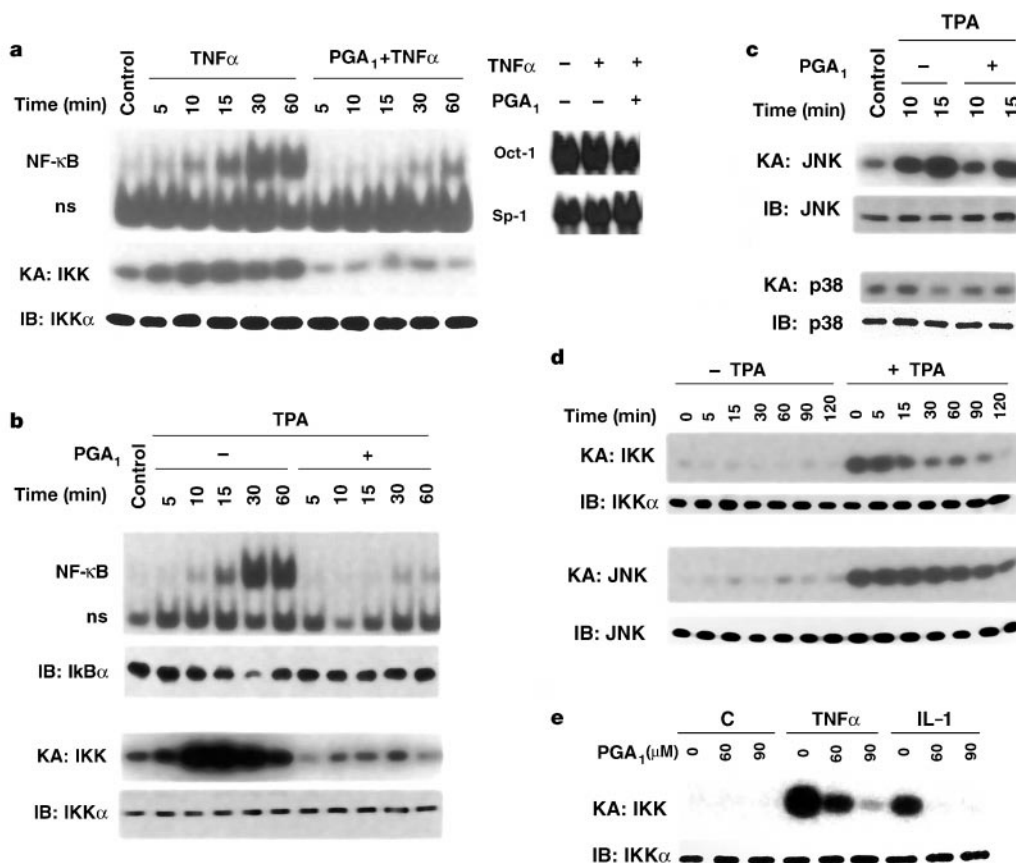


Figure 1 PGA₁ inhibits TNF α - and TPA-induced IKK and NF- κ B activities. **a**, Jurkat cells treated with 24 μ M PGA₁ or diluent for 2 h were stimulated with TNF α . At the indicated times cell lysates were prepared and assayed for NF- κ B activation by EMSA (top panel) and IKK activity by kinase assay (KA, middle panel). Positions of NF- κ B–DNA (NF- κ B) and a nonspecific protein–DNA (ns) complexes are indicated. IKK recovery was determined by immunoblotting (IB) for IKK α (bottom panel). No effects of PGA₁ (120 μ M) on Oct-1 and Sp-1 DNA-binding activities were observed (right). **b**, Jurkat cells treated with 24 μ M PGA₁ or diluent for 2 h were stimulated with TPA for the indicated times after which lysates

were prepared and assayed for NF- κ B activation (top panel), I κ B α degradation (second panel) and endogenous IKK activity (bottom panels). **c**, Samples from **b** were analysed for JNK1 and p38 activities. Recoveries of JNK and p38 were determined by immunoblotting. **d**, Jurkat cells pretreated with 24 μ M PGA₁ for the indicated times were stimulated for 15 min with TPA. Endogenous IKK and JNK activities and recoveries were determined. **e**, HeLa cells pretreated with PGA₁ were left untreated (C) or stimulated for 10 min with TNF α or IL-1. Cell lysates were assayed for endogenous IKK activity and recovery.

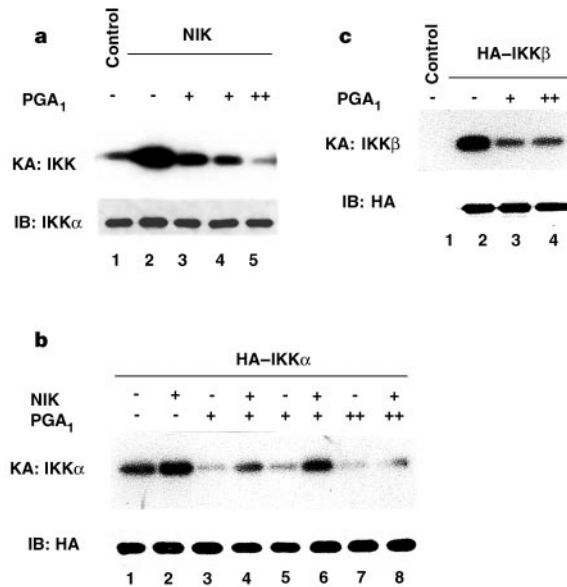


Figure 2 PGA₁ prevents NIK-induced IKK activation and inhibits IKK β activity. **a**, COS cells transfected with Xpress-tagged NIK (lanes 2–5) or 'empty' (lane 1) expression vectors were treated with PGA₁ (15 μ M in lanes 3 and 4; 30 μ M in lane 5) either 10 min (lane 3) or 18 h (lanes 4 and 5) after transfection. Cells were lysed and assayed for endogenous IKK activity and recovery 24 h after transfection. **b**, COS cells co-transfected with HA-tagged IKK α together with 'empty' (–) or Xpress-tagged NIK (+) expression vectors, and treated with PGA₁ ('+' lanes, 15 μ M; '++' lanes, 30 μ M) 10 min (lanes 3 and 4) or 18 h (lanes 5–8) after transfection. Cells were lysed and assayed for HA-tagged IKK α -associated kinase activity and recovery 24 h after transfection. **c**, COS cells were transfected with empty (control, lane 1) or HA-tagged IKK β (lanes 2–4) vectors and treated with PGA₁ ('+' lanes, 15 μ M; '++' lanes, 30 μ M) 18 h after transfection. Cells were lysed and assayed for HA-tagged IKK β kinase activity and recovery 24 h after transfection.

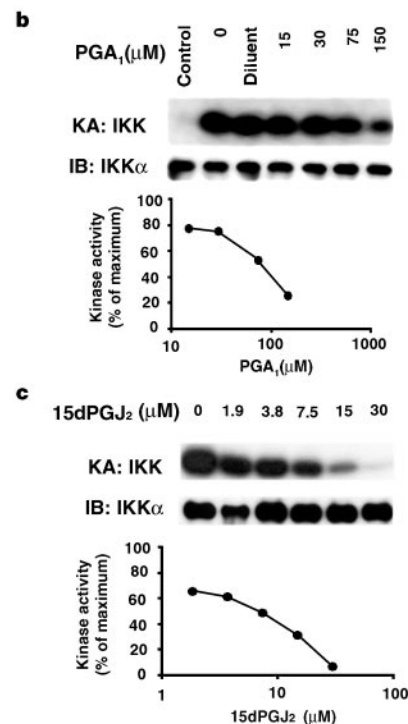
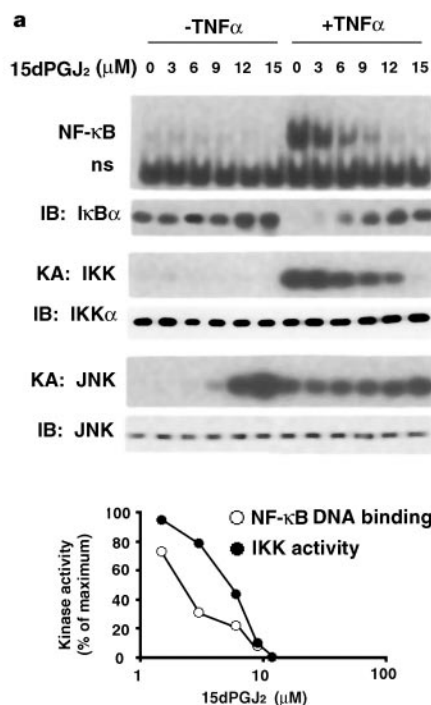


Figure 3 15dPGJ₂ inhibits IKK activity *in vivo* and *in vitro*. **a**, HeLa cells pretreated with 15dPGJ₂ for 2 h were left untreated or stimulated with TNF α for 10 min. Extracts were assayed for NF- κ B binding activity and I κ B α degradation, and for endogenous IKK and JNK1 activities (KA) and recoveries (IB). NF- κ B and IKK activities were quantified by phosphorimaging and expressed as percentage of the maximal activity achieved in the

also inhibited MEKK1-induced IKK activity in Jurkat and in COS cells (data not shown). Of the two catalytic subunits, IKK β has the major role in responding to pro-inflammatory stimuli^{10,11} and is also inhibited by high concentrations of the anti-inflammatory drug aspirin¹⁸. To determine the effect of cyPG on IKK β activity, we generated constitutively active IKK complexes, comprising mostly IKK β homodimers⁴ by overexpression of HA-tagged IKK β , in COS cells. In these cells, PGA₁ did not affect IKK β expression, as determined by immunoblotting, whereas it strongly inhibited IKK β activity (Fig. 2c).

The bioactive cyPG 15-deoxy- Δ^{12-14} -PGJ₂ (15dPGJ₂), which is physiologically formed by dehydration and isomerization of the COX metabolite PGD₂, can activate peroxisome proliferator-activated receptor- γ (PPAR- γ), a nuclear receptor that interferes with NF- κ B transcriptional activity^{8,9}. 15dPGJ₂ inhibits the production of pro-inflammatory cytokines and inducible nitric oxide synthase (iNOS) in activated monocytes by a mechanism suggested to involve PPAR- γ ^{8,9}. An anti-inflammatory effect of endogenous and exogenous 15dPGJ₂ has been shown in carrageenin-induced pleurisy in rats, which suggests that there is a physiological anti-inflammatory role for cyPGs⁷. We found that 15dPGJ₂ is a potent inhibitor of NF- κ B activation by TPA in human cells¹⁴ that express very low levels of PPAR- γ (unpublished data). We therefore examined whether 15dPGJ₂ targets IKK. In HeLa cells, which do not express PPAR- γ ⁸, 15dPGJ₂ inhibited the induction of IKK and NF- κ B activities by TNF α (Fig. 3a): the IC₅₀ for inhibition of NF- κ B DNA-binding activity was 2.25 μ M, whereas that for IKK activity was 5.08 μ M. These concentrations are comparable to those required for inhibition of cytokine synthesis^{8,9}. Intriguingly, 15dPGJ₂ stimulated JNK activity even in the absence of TNF α (Fig. 3a). 15dPGJ₂ also inhibited TNF α -induced IKK activity in human peripheral blood monocytes, but did not inhibit the DNA-binding activities of Oct1 or Sp1 (data not shown).

CyPGs also inhibited IKK activity *in vitro*. We immunoprecipitated endogenous IKK from TNF α -stimulated HeLa cells and

absence of the inhibitor. The dose–response curves represent average results from two separate experiments. **b**, **c**, HeLa cells were stimulated with TNF α for 15 min. IKK was immunoprecipitated and incubated *in vitro* with the indicated concentrations of PGA₁ (**b**), 15dPGJ₂ (**c**) or diluent for 1 h followed by kinase assay (KA) and immunoblot (IB). Kinase activity was quantified by phosphorimaging and used for plotting the dose–response curves.

determined its kinase activity in the presence of different concentrations of PGA_1 (Fig. 3b) or 15dPGJ_2 (Fig. 3c). IKK activity was inhibited in a dose-dependent fashion by both PGA_1 and 15dPGJ_2 with IC_{50} values of $79.8\ \mu\text{M}$ and $7.08\ \mu\text{M}$, respectively. No significant inhibition of either p38 or JNK1 activity was observed upon *in vitro* incubation with up to $60\ \mu\text{M}$ 15dPGJ_2 (data not shown).

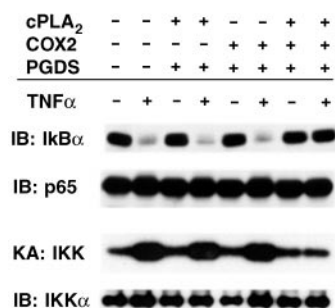


Figure 4 Inhibition of IKK activation through endogenously synthesized PGs (PGD and 15dPGJ_2). HEK 293 cells were transfected with expression vectors for cytosolic phospholipase A₂ (cPLA₂), cyclooxygenase 2 (COX2) and prostaglandin D synthase (PGDS) as indicated. After 12 h, cells were washed, incubated in low-serum (0.5%) medium for 30 h, and treated with TNF α for 15 min. I κ B α degradation and IKK activity were determined. The I κ B α immunoblot was reprobed with anti-p65 to verify equal loading of the samples.

To determine whether endogenous cyPGs are sufficient for inhibition of IKK, we transfected 293 cells with expression vectors for enzymes implicated in biosynthesis of PGD₂, which is spontaneously converted to 15dPGJ_2 by non-enzymatic dehydration^{7,19}. Combined expression of the arachidonic-acid-releasing enzyme cytosolic phospholipase A₂ (cPLA₂) with COX2 and PGD synthase (PGDS) resulted in the complete inhibition of TNF α -induced IKK activation and I κ B phosphorylation and degradation (Fig. 4). However, overexpression of each pair of enzymes (that is, cPLA₂ + PGDS and COX2 + PGDS) was insufficient for inhibition; thus, levels of cyPG that inhibit IKK can be obtained *in vivo*.

Only A- and J-type cyPGs inhibited IKK activity *in vivo* or *in vitro*; arachidonic acid and arachidonic-acid metabolites including PGB₂, PGD₂, PGE₁, PGE₂, PGF_{1 α} , thromboxane B₂ (TxB₂) and leukotriene B₄ (LTB₄) were ineffective (Fig. 5a, b). Thus, a reactive α,β -unsaturated carbonyl group in the cyclopentane ring, which renders this portion of the molecule able to form Michael adducts with cellular nucleophiles and covalently modify specific proteins^{19,20}, is essential for IKK inhibition (Fig. 5c).

The requirement for a chemically reactive cyclopentenone moiety indicated that cyPGs may inhibit IKK β through its direct modification. Many protein kinase inhibitors compete for ATP-binding sites²¹, but pre-incubation in the presence of increasing concentrations of ATP had no effect on inhibition of IKK β by 15dPGJ_2 (Fig. 6a). We also pre-incubated purified recombinant IKK β immobilized on beads with 15dPGJ_2 or diluent, and rinsed it extensively with kinase buffer lacking the inhibitor. Despite exten-

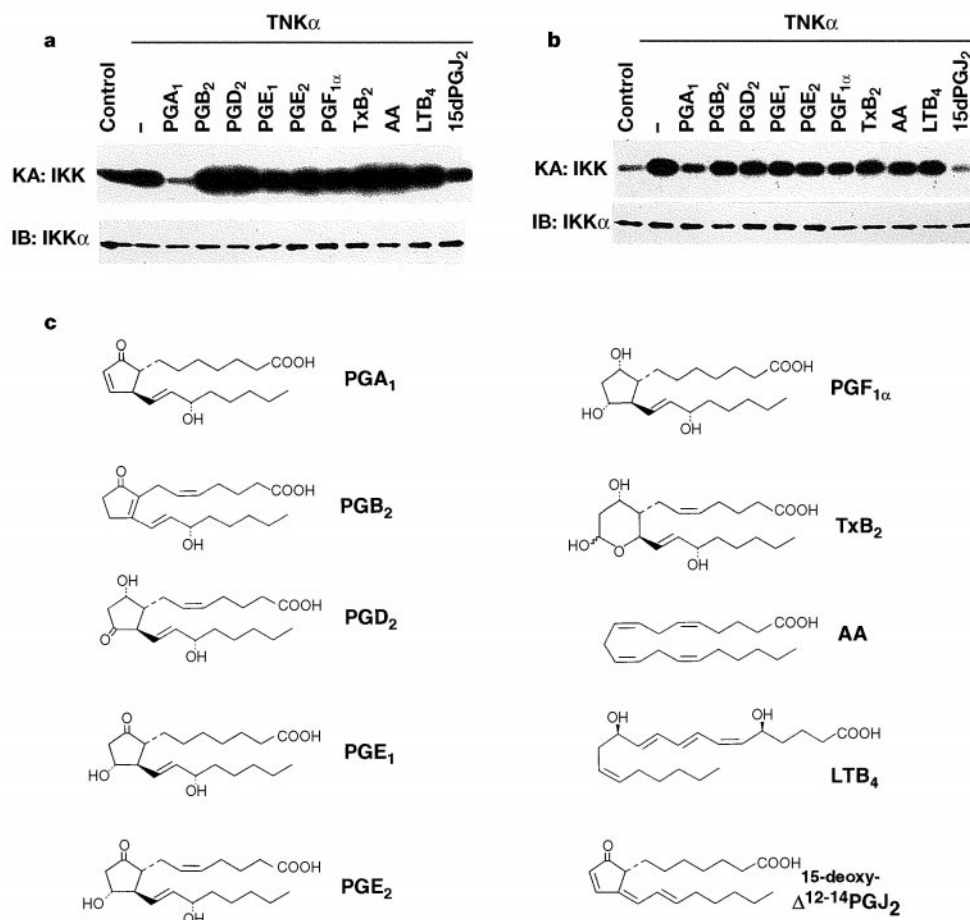


Figure 5 Effect of arachidonic acid (AA) and its metabolites on IKK activity *in vivo* and *in vitro*. **a**, HeLa cells were treated with $50\ \mu\text{M}$ AA, PGA_1 , PGB_2 , PGD_2 , PGE_1 , PGE_2 , $\text{PGF}_{1\alpha}$ or TxB_2 , or $1\ \mu\text{M}$ LTB_4 , or $5\ \mu\text{M}$ 15dPGJ_2 for 2 h in RPMI containing 10% fetal calf serum, and stimulated with TNF α for 10 min. Cell lysates were assayed for endogenous IKK activity and recovery. Control, unstimulated cells; —, treatment with diluent alone. **b**, IKK

was immunoprecipitated from TNF α -stimulated HeLa cells and incubated *in vitro* with $100\ \mu\text{M}$ AA, PGA_1 , PGB_2 , PGE_1 , PGE_2 , $\text{PGF}_{1\alpha}$ or TxB_2 or $2\ \mu\text{M}$ LTB_4 or $10\ \mu\text{M}$ 15dPGJ_2 followed by kinase assay and immunoblot analysis. Control, unstimulated cells; —, 1% ethanol. **c**, The structure of the different arachidonic-acid metabolites used in **a** and **b**.

sive washing, the sample that was pre-incubated with 15dPGJ₂ remained inhibited (data not shown). These results suggested that 15dPGJ₂ may inhibit IKK β through direct modification. Plausible targets for Michael addition reactions on proteins are cysteine residues. Both IKK β and IKK α , but not JNK1 or p38, contain a

cysteine at position 179 within their activation loop (Fig. 6b). To examine whether this cysteine is critical for sensitivity to cyPG, we replaced it with an alanine using site-directed mutagenesis. Wild-type IKK β and IKK β (C179A) were co-expressed with NIK in HeLa cells, and their sensitivity to 15dPGJ₂ was examined. Although both constructs were equally responsive to NIK, resulting in similar levels of kinase activity, the IKK β (C179A) mutant was resistant to the inhibitor (Fig. 6c). Recombinant IKK β (C179A) was also resistant to concentrations of 15dPGJ₂ that were inhibitory to wild-type IKK β (Fig. 6d) and also to PGA₁ (data not shown). We also co-transfected an NF- κ B-dependent transcriptional reporter (2XNF κ B-LUC) with an expression vector encoding either wild-type IKK β or IKK β (C179A). A NIK expression vector was included to ensure maximal IKK activation. Whereas activation of NF- κ B transcription activity through wild-type IKK β was highly sensitive to 15dPGJ₂, NF- κ B activation through IKK β (C179A) was insensitive to the inhibitor (Fig. 6e). Together, these results indicate that cyPGs inhibit IKK by direct modification of its IKK β subunit. The transfection experiments indicate that IKK β is the critical target for cyPGs in the NF- κ B activation pathway. When IKK β is rendered resistant to modification, very little inhibition of NF- κ B transcriptional activity occurs in cyPG-treated cells.

CyPGs affect cell proliferation^{19,22}, and inhibit inflammation^{7,8} and virus replication^{12,23}. Inhibition of NF- κ B activation—a factor that modulates cell proliferation, differentiation and survival, controls inflammation and is required for expression and replication of certain viruses, such as HIV-1—may account for many of these effects. Although inhibition of NF- κ B-mediated transcription by cyPG has been attributed to the activation of PPAR- γ , which interferes with NF- κ B transcriptional activity^{8,9}, such a mechanism probably requires high levels of PPAR- γ , which are not found in many cell types. In addition, potent PPAR- γ agonists, which are not cyPGs, are poor NF- κ B inhibitors^{8,24}. In particular, troglitazone, a PPAR- γ agonist, does not inhibit either IKK activity or NF- κ B activation in HeLa cells stimulated with TNF α (data not shown). Here we have described a different, PPAR- γ -independent mechanism that explains the ability of cyPG to inhibit NF- κ B. Prostaglandins are physiologically present in body fluids at picomolar-to-nanomolar concentrations^{19,25}; however, arachidonic-acid metabolism is highly increased in several pathological conditions, including hyperthermia, infection and inflammation^{5,6,26,27}, and local prostaglandin concentrations in the micromolar range have been detected at sites of acute inflammation²⁸. Elevated cyPG synthesis has been detected in late phases of inflammation⁸ and is associated with resolution of inflammation⁷. Therefore, concentrations of cyPG that are sufficient for IKK inhibition may occur locally during late phases of inflammatory responses. Moreover, as cyPGs probably inhibit IKK through covalent and irreversible modification, their effect could be cumulative. As COX2 synthesis is under NF- κ B control^{5,6}, the inhibition of IKK by cyPGs may form a negative autoregulatory loop that contributes to resolution of inflammation. Our results suggest that cyPG and possibly more potent derivatives may have therapeutic value in the treatment of inflammatory and viral diseases, as well as certain cancers²⁹ in which inhibition of NF- κ B activity may be desirable. □

Methods

Cell culture, transfection and treatments

Cells were cultured in either RPMI 1640 (Jurkat cells) or Dulbecco-modified minimum essential medium (DMEM) (HeLa or COS cells), supplemented with 10% fetal calf serum and antibiotics. Cells were stimulated with either 25 ng ml⁻¹ TPA, 20 ng ml⁻¹ recombinant human TNF α or 10 ng ml⁻¹ IL-1 β (Sigma). PGA₁, 15dPGJ₂, TxB₂, LTB₄, arachidonic acid and other PGs (Cayman Chemical Co.) were dissolved in ethanol and used as described¹⁴. Controls received equal amounts of ethanol. Transfections were carried out using Lipofectamine Plus (GIBCO) for HeLa and COS cells and DEAE-dextran (Sigma) for Jurkat cells. Xpress-tagged NIK, Xpress-tagged MEKK1, HA-tagged IKK α and HA-tagged IKK β have been described¹⁰. COX2 and cPLA₂ expression vectors (gifts from L. Feng) have been described³⁰. PGD synthase was cloned from an HT29 colon carcinoma cell line

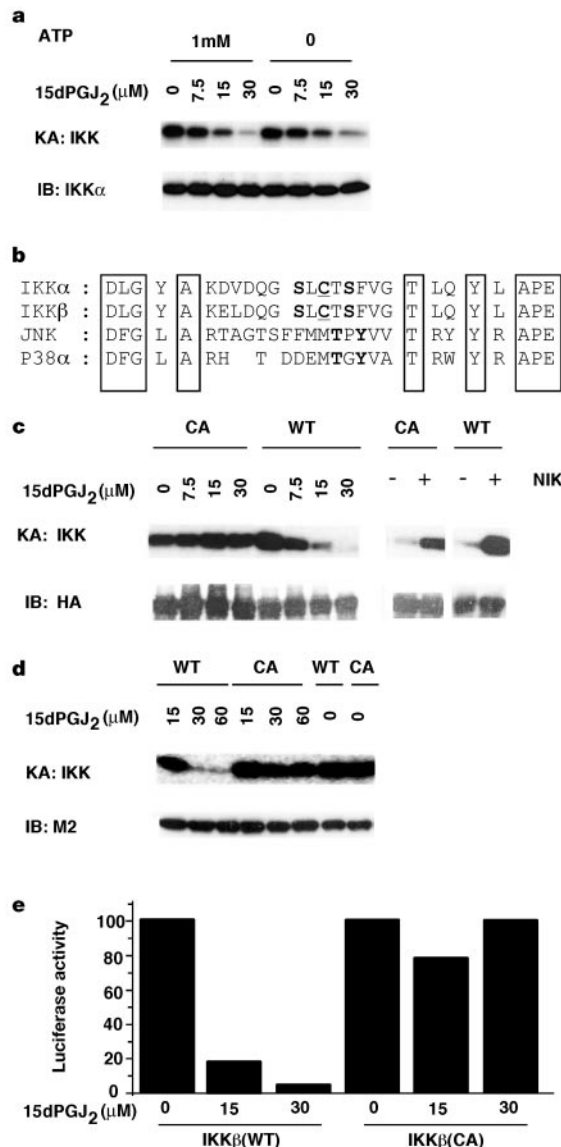


Figure 6 15dPGJ₂ is a direct inhibitor of IKK β , whose sensitivity to inhibition is mediated by a cysteine in the activation loop. **a**, 15dPGJ₂ is not a competitive inhibitor of ATP binding. IKK was immunoprecipitated from TNF α -stimulated HeLa cells and incubated with 15dPGJ₂ (0–30 μ M) with or without ATP in lysis buffer at 4 °C for 30 min. The pellets were rinsed and kinase activity measured. **b**, Sequence alignment of the activation loops of IKK α , IKK β , JNK1 and p38 α . Phospho-acceptor sites critical for kinase activation are in bold; cysteine residues of IKK α and IKK β are underlined. **c**, HeLa cells were transiently transfected with expression vectors for wild-type IKK β (WT) or IKK β (C179A) (CA) with or without a NIK expression vector. After 16 h, NIK-transfected cells were treated for 2 h with 15dPGJ₂ (left panel). Kinase activity (KA) and expression (IB) were subsequently determined. **d**, Sf9 cells infected with FLAG-tagged wild-type IKK β and IKK β (C179A) baculoviruses were incubated with various concentrations of 15dPGJ₂ for 2 h at 28 °C, and kinase activity and expression determined. **e**, HeLa cells were transiently transfected with a 2XNF- κ B-LUC reporter, a NIK expression vector and expression vectors for wild-type IKK β (IKK β (WT)) or IKK β (C179A) (IKK β (CA)). Cells were treated with 15dPGJ₂ at 3 h after transfection, and luciferase activity was determined of 24 h on duplicate samples. In each case the maximal activity achieved in the absence of inhibitor was given an arbitrary value of 100% and the other values were calculated relative to that.

complementary DNA library by PCR and subcloned in pCDNA3 vector. Cells were lysed in buffer B (50 mM Tris-HCl pH 7.5, 400 mM NaCl, 1 mM EDTA, 1 mM EGTA, 1% Triton, 0.5% NP-40 and 10% glycerol supplemented with 20 mM β -glycerophosphate, 19 mM PNPP, 500 mM Na₂VO₄, 1 mM PMSF, 20 μ g ml⁻¹ aprotinin, 2.5 μ g ml⁻¹ leupeptin, 8.3 μ g ml⁻¹ bestatin, 1.7 μ g ml⁻¹ pepstatin and 2 mM dithiothreitol.

Electrophoretic mobility-shift assay (EMSA)

Aliquots of total extracts (15 μ g protein) in buffer B were incubated with a ³²P-labelled NF- κ B DNA, followed by analysis of DNA-binding activity by EMSA on 4% polyacrylamide gels¹⁴. The specificity of protein-DNA complexes was verified by supershift with polyclonal antibodies specific for p65 (Rel A).

Kinase assay, immunoprecipitation and immunoblotting

Cell lysates were incubated with anti-IKK α (Pharmingen), anti-HA (12CA5; Pharmingen), anti-JNK1 (333.8; Pharmingen) or anti-p38 (New England Biolabs) for 2 h, and 15 μ l protein-A-Sepharose (Sigma) was added for 1 h. After extensive washing, kinase assays were done as described³. Endogenous IKK, HA-tagged IKK α , HA-tagged IKK β , endogenous JNK1 and p38 activities were determined with GST-IkB α (1-54), GST-c-Jun(1-79) and GST-ATF2 as substrates. Recombinant IKK β was expressed in Sf9 cells and purified as described⁴. Arachidonic acid, PGs, LTB₄ and TxB₂ were added to washed immunoprecipitates of recombinant IKK β in kinase buffer in a final volume of 500 μ l. After 1.5 h at 4 °C, the immunoprecipitates were washed and the kinase activity was determined. Western blot analysis was done as described¹⁴.

Received 2 August; accepted 17 November 1999.

- Baldwin, A. S. The NF- κ B and I κ B proteins: new discoveries and insights. *Annu. Rev. Immunol.* **14**, 649-683 (1996).
- Barnes, P. J. & Karin, M. Nuclear factor- κ B—a pivotal transcription factor in chronic inflammatory diseases. *New Engl. J. Med.* **336**, 1066-1071 (1997).
- DiDonato, J. A., Hayakawa, M., Rothwarf, D. M., Zandi, E. & Karin, M. A cytokine-responsive I κ B kinase that activates the transcription factor NF- κ B. *Nature* **388**, 548-554 (1997).
- Zandi, E., Chen, Y. & Karin, M. Direct phosphorylation of I κ B by IKK α and IKK β : discrimination between free and NF- κ B-bound substrate. *Science* **281**, 1360-1363 (1998).
- Herschman, H. R. *et al.* Function and regulation of prostaglandin synthase-2. *Adv. Exp. Med. Biol.* **407**, 61-66 (1997).
- Newton, R., Kintert, L. M. E., Bergmann, M., Adcock, I. M. & Barnes, P. I. Evidence for involvement of NF- κ B in the transcriptional control of COX-2 gene expression by IL-1 β . *Biochem. Biophys. Res. Comm.* **237**, 28-32 (1997).
- Gilroy, D. W. *et al.* Inducible cyclooxygenase may have anti-inflammatory properties. *Nature Med.* **5**, 698-701 (1999).
- Ricote, M., Li, A. C., Willson, T. M., Kelly, C. J. & Glass, C. K. The peroxisome proliferator-activated receptor- γ is a negative regulator of macrophage activation. *Nature* **391**, 79-82 (1998).
- Jiang, C., Ting, A. T. & Seed, B. PPAR- γ agonists inhibit production of monocyte inflammatory cytokines. *Nature* **391**, 82-86 (1998).
- Delhase, M., Hayakawa, M., Chen, Y. & Karin, M. Positive and negative regulation of I κ B kinase activity through IKK β subunit phosphorylation. *Science* **284**, 309-313 (1999).
- Tanaka, M. *et al.* Embryonic lethality, liver degeneration and impaired NF- κ B activation in IKK- β deficient mice. *Immunity* **10**, 421-429 (1999).
- Santoro, M. G. Antiviral activity of cyclopentenone prostanoids. *Trends Microbiol.* **5**, 276-281 (1997).
- Rozera, C. *et al.* Inhibition of HIV-1 replication by cyclopentenone prostaglandins in acutely infected human cells. *J. Clin. Invest.* **97**, 1795-1803 (1996).
- Rossi, A., Elia, G. & Santoro, M. G. Inhibition of nuclear factor κ B by prostaglandin A₁: an effect associated with heat shock transcription factor activation. *Proc. Natl Acad. Sci. USA* **94**, 746-750 (1997).
- Rothwarf, D. M., Zandi, E., Natoli, G. & Karin, M. IKK- γ is an essential regulatory subunit of the I κ B kinase complex. *Nature* **395**, 297-300 (1998).
- Yamaoka, S. *et al.* Complement cloning of NEMO, a component of the I κ B kinase complex essential for NF- κ B activation. *Cell* **93**, 1231-1240 (1998).
- Karin, M. & Delhase, M. JNK or IKK, AP-1 or NF- κ B, which are the targets for MEK kinase 1 action? *Proc. Natl Acad. Sci. USA* **95**, 9067-9069 (1998).
- Yin, M., Yamamoto, Y. & Gaynor, R. B. The anti-inflammatory agents aspirin and salicylate inhibit the activity of I κ B kinase- β . *Nature* **398**, 77-80 (1998).
- Fukushima, M. Prostaglandin J2-anti-tumor and anti-viral activities and the mechanisms involved. *Eicosanoids* **3**, 189-199 (1990).
- Rossi, A., Elia, G. & Santoro, M. G. 2-Cyclopenten-1-one, a new inducer of heat shock protein 70 with antiviral activity. *J. Biol. Chem.* **271**, 32192-32196 (1996).
- Gray, N. S. *et al.* Exploiting chemical libraries, structure and genomics in the search for kinase inhibitors. *Science* **281**, 533-537 (1998).
- Santoro, M. G., Garaci, E. & Amici, C. Prostaglandins with antiproliferative activity induce the synthesis of a heat shock protein in human cells. *Proc. Natl Acad. Sci. USA* **86**, 8407-8411 (1989).
- Santoro, M. G., Benedetto, A., Carruba, G., Garaci, E. & Jaffe, B. M. Prostaglandin A compounds as antiviral agents. *Science* **209**, 1032-1034 (1980).
- Stael, B. *et al.* Activation of human aorta smooth-muscle cells is inhibited by PPAR α but not by PPAR γ activators. *Nature* **393**, 790-793 (1998).
- Serhan, C. N., Haegström, J. Z. & Leslie, C. C. Lipid mediator networks in cell signaling: update and impact of cytokines. *FASEB J.* **10**, 1147-1158 (1996).
- Calderwood, S. K., Bornstein, B., Farnum, E. K. & Stevenson, M. A. Heat shock stimulates the release of arachidonic acid and the synthesis of prostaglandins and leukotriene B₄ in mammalian cells. *J. Cell Physiol.* **141**, 325-333 (1989).
- Leung, D. Y. *et al.* Increased *in vitro* bone resorption by monocytes in the hyper-immunoglobulin E syndrome. *J. Immunol.* **140**, 84-88 (1988).
- Offenbacher, S., Odle, B. M. & Van Dyke, T. E. The use of crevicular fluid prostaglandin E₂ levels as a predictor of periodontal attachment loss. *J. Periodontol. Res.* **21**, 101-112 (1986).

- Wang, C. Y., Mayo, M. W. & Baldwin, A. S. TNF and cancer therapy-induced apoptosis: potentiation by inhibition of NF- κ B. *Science* **274**, 786-787 (1998).
- Hirose, S. *et al.* Expression and localization of cyclooxygenase isoforms and cytosolic phospholipase A₂ in anti-Thy-1 glomerulonephritis. *J. Am. Soc. Nephrol.* **9**, 408-416 (1998).

Acknowledgements

We thank L. Feng for useful suggestions and for cPLA₂ and COX2 expression vectors. This work was supported by grants from the Italian Ministry of Public Health, II National AIDS Research Project (1999), the CNR Target Project on Biotechnology (G.S.), the NIH and the California Cancer Research Program (M.K.). M.K. is the Frank and Else Schilling American Cancer Society Research Professor. P.K., G.N. and Y.C. were supported by postdoctoral fellowships from the Wellcome Trust, Damon-Runyon Walter-Winchell Cancer Research Fund and the Tobacco Related Disease Research Program, respectively.

Correspondence and requests for material should be addressed to M.K. (e-mail: karinoffice@ucsd.edu).

Nuclear pore complexes in the organization of silent telomeric chromatin

Vincent Galy*, Jean-Christophe Olivo-Marín†, Harry Scherthan‡, Valerie Doye§, Nadia Rascalou* & Ulf Nehrbass*

* Laboratoire de Biologie Cellulaire du Noyau, Institut Pasteur, CNRS URA 1773, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France

† Laboratoire d'Analyse d'Images Quantitative, Institut Pasteur, CNRS URA 1947, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France

‡ Abt. Humanbiologie and Abt. Zellbiologie, der Universität, Postf. 3049, D-67653 Kaiserslautern, Germany

§ Institut Curie, CNRS, UMR144, 22 rue Lhomond, 75248 Paris Cedex 05, France

The functional regulation of chromatin is closely related to its spatial organization within the nucleus. In yeast, perinuclear chromatin domains constitute areas of transcriptional repression¹⁻³. These 'silent' domains are defined by the presence of perinuclear telomere clusters⁴. The only protein found to be involved in the peripheral localization of telomeres is Yku70/Yku80 (ref. 5). This conserved heterodimer⁶ can bind telomeres⁷ and functions in both repair of DNA double-strand breaks⁸⁻¹¹ and telomere maintenance^{7,12-15}. These findings, however, do not address the underlying structural basis of perinuclear silent domains. Here we show that nuclear-pore-complex extensions formed by the conserved TPR^{16,17} homologues Mlp1 and Mlp2^{18,19} are responsible for the structural and functional organization of perinuclear chromatin. Loss of MLP2 results in a severe deficiency in the repair of double-strand breaks. Furthermore, double deletion of MLP1 and MLP2 disrupts the clustering of perinuclear telomeres and releases telomeric gene repression. These effects are probably mediated through the interaction with Yku70. Mlp2 physically tethers Yku70 to the nuclear periphery, thus forming a link between chromatin and the nuclear envelope. We show, moreover, that this structural link is docked to nuclear-pore complexes through a cleavable nucleoporin, Nup145²⁰. We propose that, through these interactions, nuclear-pore complexes organize a nuclear subdomain that is intimately involved in the regulation of chromatin metabolism.

Mlp1 and Mlp2 are putative coiled-coil proteins that are located at the interface between the nuclear envelope and the nuclear interior¹⁹. Sequence homology and similar behaviour in biochemical fractionation suggest that the two proteins have overlapping functions¹⁹. Although double mutants did not have an obvious

defect in nuclear transport (data not shown), they formed irregular colonies with deep invaginations typical of DNA repair mutants (data not shown). Single or double mutants showed only minor sensitivity to ultraviolet damage¹⁸ and showed no increase in generation time (data not shown). We tested for DNA double-strand break (DSB) repair using the radiomimetic antibiotic bleomycin, and found that deletion of *MLP2* leads to a marked decrease in repair efficiency (Fig. 1a). A defect in DSB repair in *MLP1* deletion mutants became apparent only at high bleomycin concentrations (Fig. 1a). Deletion of both *MLP1* and *MLP2* gave the same phenotype as loss of *MLP2* (Fig. 1a). When compared with known DSB repair mutants, the extent of DNA repair deficiency in $\Delta mlp2$ cells was equivalent to deletion of *YKU70* (Fig. 1b) and weaker than that displayed by a *RAD50* deletion mutant (data not shown).

To analyse whether deletion of *MLP2* has a direct effect on Yku70, we studied the subnuclear distribution of a Yku70–myc fusion protein in indirect immunofluorescence. Wild-type cells showed a peripheral pattern and absence of signal from the nuclear interior (Fig. 2a, upper panel), reminiscent of Yku80 staining²¹. In contrast, cells lacking *MLP2* showed a dispersed nucleoplasmic Yku70–myc signal (Fig. 2a, lower panel). To assess the Yku70–myc redistribution quantitatively, we used a dedicated image analysis program to determine the distance of signal foci from the nuclear centre. The radial distribution of foci in wild-type cells had a peak corresponding to a predominantly perinuclear localization (Fig. 2b), whereas the mutant had fewer Yku70–myc signal foci, which were uniformly distributed throughout the nuclear plane (Fig. 2b). These data indicate that Mlp2 may be required for the peripheral localization of Yku70.

We carried out co-immunoprecipitation experiments to address

whether Mlp2 is directly involved in localizing Yku70 to the nuclear periphery. We recombined a *Staphylococcus aureus* protein A tag to the 3' end of *MLP2* (Mlp2–A) in a strain that contained a myc-tagged version of *YKU70* (Yku70–myc). Extracts precipitated for Mlp2–A by IgG-sepharose showed specific co-precipitation of Yku70–myc (Fig. 2c). In contrast, Rap1 (ref. 22), another telomere-binding factor, was absent from these fractions (Fig. 2c), showing a physical interaction between Mlp2 and Yku70. To stress the functional significance of this interaction further, we studied the nuclear matrix partitioning of Yku70. Either wild-type or $\Delta mlp2$ cells expressing Yku70–myc were extracted under conditions in which more than 70% of both Mlp1 and Mlp2 partition with the insoluble matrix fraction (data not shown). Figure 2d shows that lack of Mlp2 resulted in a threefold decrease in Yku70–myc partitioning, with less than 20% of the protein remaining bound to the matrix. The fractionation behaviour of yeast fibrillarin²³, a nuclear matrix protein, was unaffected by the lack of Mlp2, as was Rap1, which did not fractionate into the nuclear matrix (Fig. 2d). Binding of Yku70–myc to the insoluble nuclear matrix thus depends on the presence of Mlp2. Together, these data imply a physical link between both proteins and show that Mlp2 tethers Yku70 to its perinuclear localization.

DNA DSBs cause a transient release of the Yku complex into the

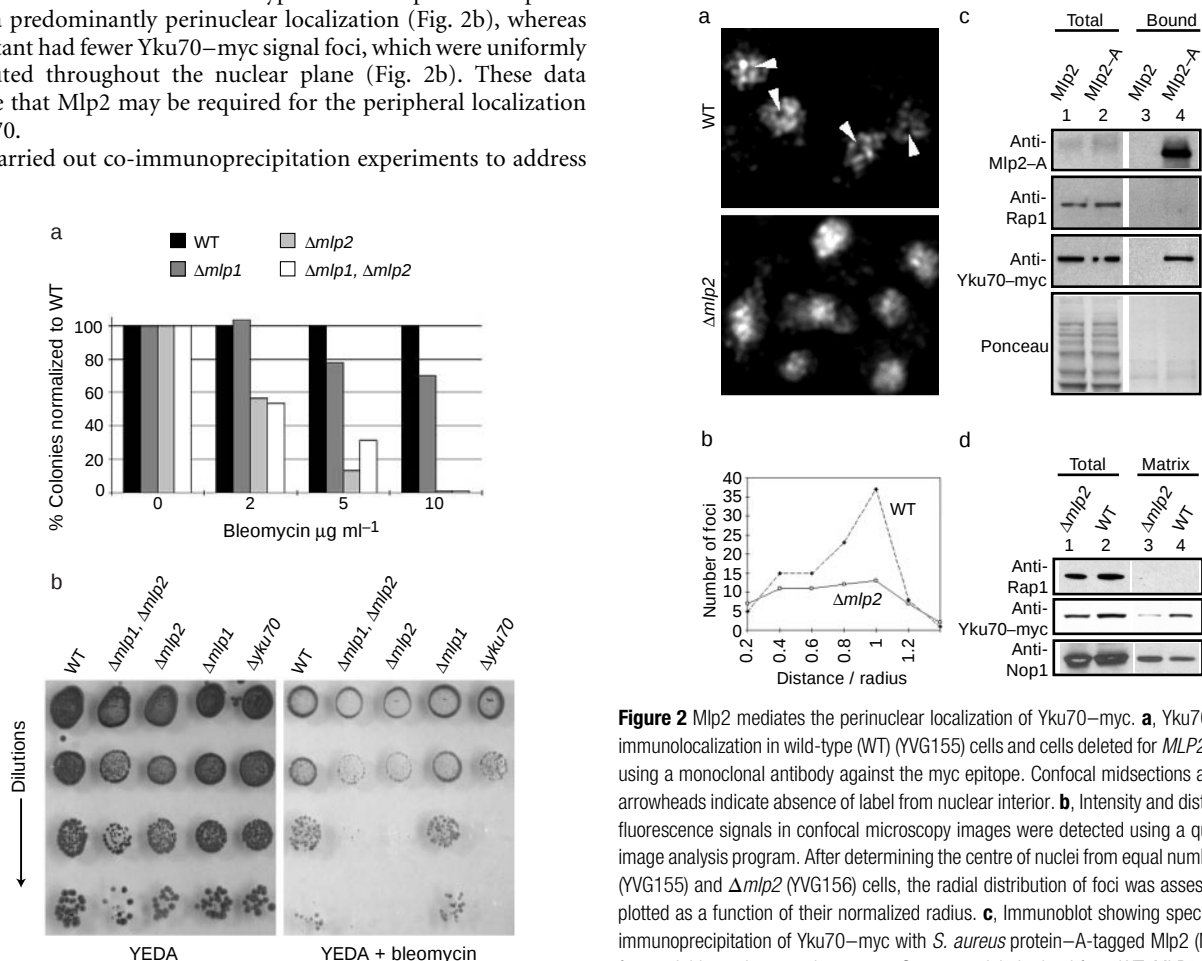


Figure 1 Deletion of *MLP2* leads to impairment of DNA DSB repair. **a**, Relative quantification of bleomycin sensitivity of $\Delta mlp1$ (YVG1), $\Delta mlp2$ (YVG3) and $\Delta mlp1, \Delta mlp2$ (YVG12) cells compared with wild-type (WT) cells (BMA64). The number of surviving WT colonies at each concentration step was normalized to 100%. **b**, Drop test on YEDA²⁷ medium in the absence or presence of bleomycin (5 $\mu\text{g ml}^{-1}$), showing bleomycin sensitivity of WT (BMA64), $\Delta mlp1, \Delta mlp2$ (YVG12), $\Delta mlp2$ (YVG3), $\Delta mlp1$ (YVG1) and $\Delta yku70$ (YVG159) strains.

Figure 2 Mlp2 mediates the perinuclear localization of Yku70–myc. **a**, Yku70–myc immunolocalization in wild-type (WT) (YVG155) cells and cells deleted for *MLP2* (YVG156) using a monoclonal antibody against the myc epitope. Confocal midsections are shown; arrowheads indicate absence of label from nuclear interior. **b**, Intensity and distribution of fluorescence signals in confocal microscopy images were detected using a quantitative image analysis program. After determining the centre of nuclei from equal numbers of WT (YVG155) and $\Delta mlp2$ (YVG156) cells, the radial distribution of foci was assessed and plotted as a function of their normalized radius. **c**, Immunoblot showing specific co-immunoprecipitation of Yku70–myc with *S. aureus* protein A-tagged Mlp2 (Mlp2–A) from soluble nuclear matrix extracts. Start material obtained from WT *MLP2* (YVG155, lane 1) and *MLP2*–A (YVG189, lane 2) cells and equivalent amounts of specific co-immunoprecipitate (800 equivalents of start material) from WT *MLP2* (lane 3) and *MLP2*–A (lane 4) were analysed. Mlp2–A labelling in the start material is at the limit of detection. **d**, Immunoblot of total lysate (lane 1 and 2) or insoluble nuclear matrix extracts (lanes 2 and 3) prepared from WT cells (YVG155, lanes 2 and 4) or cells lacking *MLP2* (YVG156, lanes 1 and 3), both containing Yku70–myc. Equivalent amounts of total lysate and matrix fractions were loaded.

nucleoplasm²¹ where, amongst other functions, the complex is required for G2/M adaptation²⁴. A lack of functional Yku70 in the absence of repair prevents cells with severed chromosome ends from resuming growth, thus leading to a permanent G2/M arrest²⁴. We found that deletion of *MLP2* in test strain JKM179¹⁰ had no effect on

G2/M adaptation (data not shown). Thus, the λ protein is unlikely to have an overall effect on Yku70. To analyse whether Mlp proteins predominantly affect Yku70 function at the nuclear periphery, we tested for Mlp involvement in telomere maintenance. Deletions of *YKU70* or *YKU80* are the only mutations known to lead to telomere mislocalization⁵. When we performed fluorescence *in situ* hybridization (FISH) in $\Delta mlp1/\Delta mlp2$ cells using a pan-telomeric X and Y' repeat probe²⁵, we found a clearly altered spatial distribution of telomeric signals (Fig. 3a, b). Quantitative image analysis determined a 42% increase in the average number of telomeric foci upon *MLP1/MLP2* deletion (Fig. 3c), similar to the effect caused by *YKU70* deletion⁵. When plotted for radial distribution (Fig. 3d), foci in wild-type cells show a peak at the maximum relative radius corresponding to the nuclear periphery, whereas mutant cells show a significant reduction in the number of peripheral foci and have a second peak at a shorter radius (Fig. 3d). This change reflects the re-localization of telomeric foci from the nuclear periphery towards the centre of the nucleus. The absence of Mlp proteins, therefore, leads to a significant mislocalization of telomeres and a disintegration of telomeric clusters. This effect is concomitant with de-repression of telomeric silencing, as assayed in a yeast strain containing a *URA3* reporter gene in the telomeric region of chromosome VII. Reporter cells lacking both *MLP1* and *MLP2* were markedly de-repressed, with the effect being ~10-fold lower than that observed for *YKU70* deletion (Fig. 3e). As the high local concentration of silencing factors in telomere clusters may have a cooperative effect on silencing at subtelomeric sites, the de-repression in $\Delta mlp1/\Delta mlp2$ cells could be due to disintegration of telomeric clusters. Alternatively, the nucleoplasmic dispersion of Yku70 may diminish the amount of Yku70 and associated Sir proteins available at telomeric sites²¹. Together, these results show that nucleoplasmic Mlp extensions are required for the perinuclear localization of telomeres and for efficient transcriptional repression of telomeric genes.

To investigate whether this Mlp-dependent organization of the perinuclear region is eventually implemented by nuclear pore complexes (NPCs), we tried to identify the potential docking site of Mlp proteins at the NPC. We used a synthetic lethality screen which has previously proven effective in identifying intra-NPC interactions. The $\Delta mlp1$ synthetic lethality (SL) screen identified several specific synthetic lethal clones, which fell into three complementation groups. One synthetic lethal clone, SL α 2, revealed intranuclear mislocalization of a Mlp1-GFP (green fluorescent protein) fusion protein, whereas NPCs retained their normal distribution pattern (Fig. 4a). SL α 2 synthetic lethality could be rescued by a plasmid carrying a full-length copy of *NUPI45* (Fig. 4b), an essential nucleoporin that is involved in pore clustering and RNA export²⁰, and is unusual in that it can undergo autocatalytic cleavage in a central domain²⁰. To probe independently for a functional interaction between *NUPI45* and *MLP1*, we generated a truncation mutant of *NUPI45* by recombining a stop codon into the internal cleavage site²⁰. Analysis of the subnuclear localization of Mlp proteins in the resulting *nup145 Δ C* cells revealed nucleoplasmic mislocalization reminiscent of SL α 2 (Fig. 4c). These results show that NPC docking of Mlp1 and Mlp2 requires Nup145. This prompted us to analyse whether the NPC-dependent mislocalization of Mlps in *nup145 Δ C* cells would reproduce the functional defects of *MLP* deletion. We tested for the effect of *nup145 Δ C* on telomere position effect (TPE) and on DSB repair. Figure 4d shows that the *nup145 Δ C* truncation resulted in de-repression of a subtelomeric *URA3* reporter gene. Growth of *nup145 Δ C* cells on bleomycin-containing medium revealed a significant DSB repair defect (Fig. 4e). Finally, a functional GFP-Rap1 fusion protein showed a marked mislocalization of telomeres in *nup145 Δ C* cells (Fig. 4f), as well as a striking increase in the number of telomeric foci (Fig. 4g).

Our results indicate that there may be a physical link between

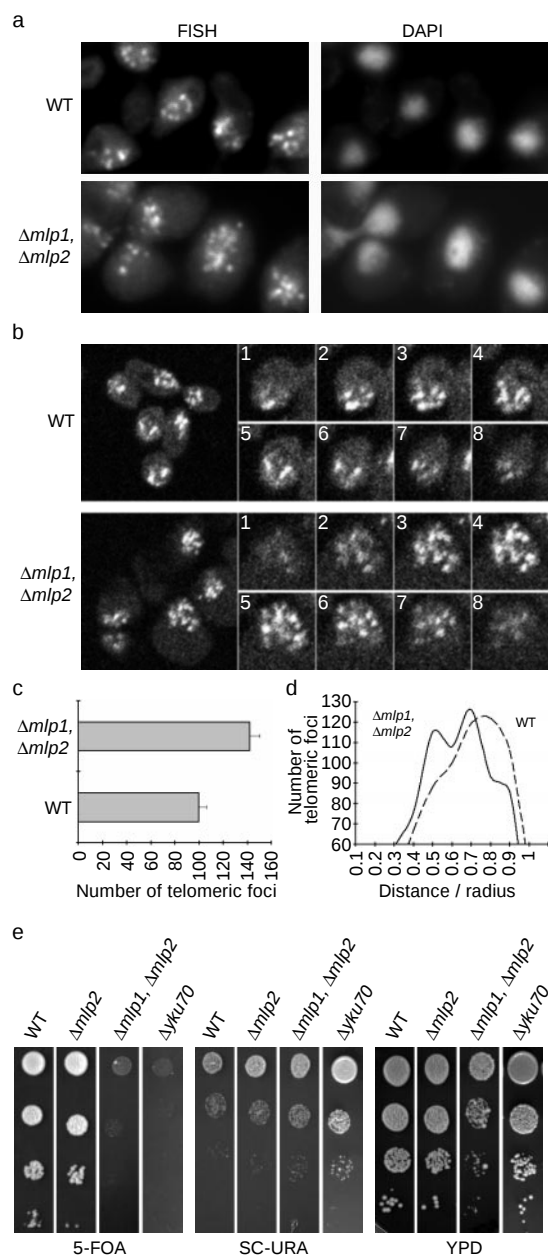


Figure 3 Double deletion of *MLP1* and *MLP2* releases telomeres from their perinuclear localization and leads to de-repression of telomeric silencing. **a**, **b**, Fluorescence *in situ* hybridization (FISH) was used to visualize subnuclear telomere distribution on structurally preserved nuclei by conventional fluorescence (**a**) or confocal microscopy (**b**). Figures show unique and representative fields of view and eight confocal sections through typical wild-type (WT) (BMA64, a/a) or mutant (YVG160, a/a) cells in 0.3 μ m steps (**b**). **c**, Telomeric foci in wild-type and mutant cells were automatically detected and counted by a dedicated program. The plot shows the result of the analysis of 100 cells, normalized to 100% for wild-type cells. **d**, The relative distributions of telomeres in the nuclei were determined by computing the normalized distance of each telomere to the centre of its corresponding nucleus. Telomere distributions in the mutant and wild type are statistically different (χ^2 -test, $P < 0.02$). **e**, Telomeric silencing was determined using WT (YG59), $\Delta mlp2$ (YVG120), $\Delta mlp1, \Delta mlp2$ (YVG127) and $\Delta yku70$ (YVG185) cells containing a *URA3* gene in the subtelomeric region of chromosome VII. Cells were grown on 5-FOA, SC-URA and YPD plates.

NPCs and telomeres. The involvement of NPCs in the subnuclear localization and functional organization of chromatin would imply that NPCs have a role beyond mediating nuclear-translocation processes. Rather than reflecting two independent commitments, both aspects of NPC function might be closely interrelated, as has previously been suggested²⁶. Moreover, our data provide evidence that Mlp1/Mlp2 and Yku70 share a significant functional overlap over a rather diverse set of metabolic processes. Together with the physical interaction between Mlp2 and Yku70, these findings lead us to conclude that Mlp proteins mediate their function in chromatin organization through Yku70. Conversely, Mlp-dependent perinuclear localization appears to be a prerequisite for several basic functions of Yku70. We propose a model in which NPC extensions formed by Mlp proteins tether chromosome ends to the nuclear periphery through interaction with Yku70, thus constituting perinuclear silent chromatin domains. The structural integrity of these topological compartments appears to be functionally relevant for TPE and DNA DSB repair, suggesting that certain aspects of repair occur in silent chromatin domains. It will be intriguing to explore whether this possible role of NPCs as anchor points for subnuclear

chromatin organization has been conserved during eukaryotic evolution. □

Methods

Strains, plasmids and FISH

For details about plasmid and strain constructions, as well as genotypes of strains used, see Supplementary Information. FISH analysis for detection of Y' telomere repeats was done as described²⁵.

Bleomycin sensitivity assay and microscopy

The bleomycin sensitivity assay was done as described²⁷. Immunofluorescence microscopy was done as described²⁸. Cells were examined on a Leica DMRXA fluorescence microscope and image acquisition was done with an Hamamatsu cooled CCD camera.

Nuclear matrix preparation and co-immunoprecipitation

A subcellular fractionation protocol was adapted from ref. 29. Immunodetections were performed with a monoclonal antibody 9E10 to detect Hdf1-myc and an anti-Rap1 goat polyclonal IgG (sc-6662, Santa Cruz Biotechnology) to detect Rap1. Nop1 was detected using a polyclonal rabbit antiserum as described²³. Co-immunoprecipitation was carried out on soluble fractions from nuclear matrix preparations using IgG-Sepharose (Pharmacia). An anti-protein-A fractionated rabbit serum was used to detect Mlp2-A (P-3775, Sigma).

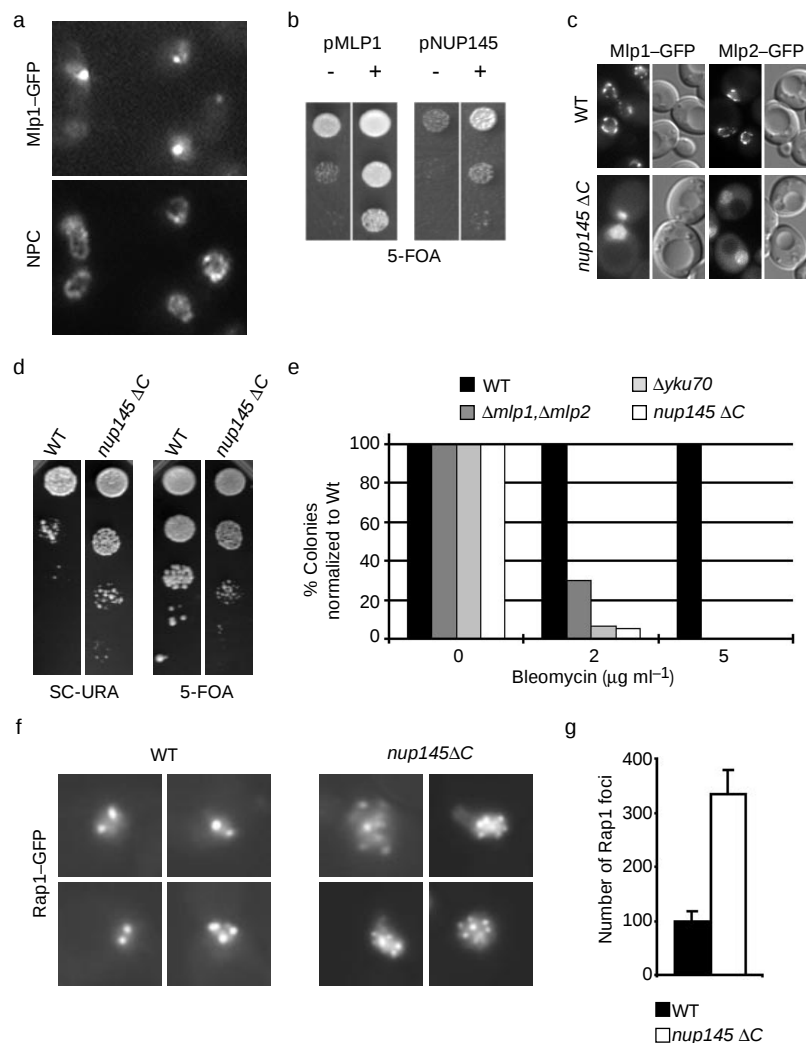


Figure 4 Mlp1 and Mlp2 are docked to the nuclear pore complex (NPC) through the cleavable nucleoporin Nup145. **a**, The SLα2 strain was transformed by pRS315MCP1-GFP and the subcellular localization of Mlp1-GFP was revealed using a polyclonal antibody against GFP and comparison with the distribution of NPCs localized using mAb414. **b**, Growth of SLα2 on 5-FOA-containing medium after complementation by a plasmid carrying either *MLP1*, *NUP145* or no insert (left columns). **c**, Mlp1-GFP and Mlp2-GFP localization in living cells grown at 25 °C and shifted to 30 °C for 4 h before GFP fluorescence analysis. **d**, Telomeric silencing was determined using wild-type (WT)

(YVG59) and *nup145ΔC-myc* (YVG186) cells containing a *URA3* gene in the subtelomeric region of chromosome VII grown on SC-URA- or 5-FOA-containing medium at 25 °C. **e**, Relative quantification of bleomycin sensitivity of *Δmlp1, Δmlp2* (YVG12), *Δyku70* (YVG159) and *nup145ΔC-GFP* (YVG179), as compared to WT cells. The number of surviving wild-type colonies at each concentration step was normalized to 100%. **f**, GFP-Rap1 localization in living WT (YVG165) and *nup145ΔC-pA* (YVG170) cells grown at 25 °C. **g**, GFP-Rap1 foci were counted and the plot shows the result of the analysis of 90 cells normalized to 100% for the WT.

Quantitative image analysis

To evaluate the distribution of telomeres in the nuclei, we devised a quantification method based on counting the telomeres and the measurement of the distance from the telomeres to the centre of the nucleus. A dedicated program was developed to detect, count and measure nuclei and telomeres from the fluorescence images. The telomere detection and counting was carried out automatically using a multiresolution algorithm based on selectively filtering an undecomposed wavelet decomposition of the image through the use of wavelet coefficient thresholding and correlation³⁰. After nuclei are determined by thresholding, a logical AND operation is applied between the images to attribute each focus to its containing nucleus. The program then computes the normalized distance d of each focus to the centre of its corresponding nucleus, defined as d/R with values $0 \leq d \leq 1$, where R is the radius of the nucleus. The values obtained for d were accumulated from over 100 cells and grouped into 10 distance categories of 0.1 step. Statistical significance was assessed by a χ^2 -test.

Telomeric silencing assay

The effect of *MLP1* and *MLP2* deletions on telomeric silencing was analysed using strain YG59 (provided by D. Shore) containing a *URA3* gene inserted into the subtelomeric *ADH4* site of chromosome VII. YG59 cells were deleted for *Yku70*, *MLP1* or *MLP2*, or were double deleted for *MLP2* and *MLP1*. The effect of the carboxy-terminal deletion of *NUP145* on the telomeric silencing was also analysed. For each strain, drops of 10-fold serial dilutions cultures were put on 5-FOA, SC-URA and YPD plates and incubated at either 30 °C or 25 °C.

Synthetic lethality screening

The screening for mutants that were synthetically lethal with $\Delta mlp1$ was carried out in an *ade3⁻ura3⁻*, $\Delta mlp1$ -strain(YVG19) carrying wild-type *MLP1* on a *URA3-ADE3* plasmid. Cells were mutagenized using ultraviolet at an intensity that left 10% surviving. From 43,000 mutagenized clones, five non-sectored and 5-FOA sensitive clones were isolated. One of these mutants was complemented by *NUP145* (plasmid pUN100-NUP145FL, provided by E. Fabre). The others remain to be characterized.

Received 7 September; accepted 16 November 1999.

- Maillet, L. *et al.* Evidence for silencing compartments within the yeast nucleus: a role for telomere proximity and Sir protein concentration in silencer-mediated repression. *Genes Dev.* **10**, 1796–1811 (1996).
- Andrulis, E. D., Neiman, A. M., Zappulla, D. C. & Sternglanz, R. Perinuclear localization of chromatin facilitates transcriptional silencing. *Nature* **394**, 592–595 (1998); erratum, *ibid* **395**, 525 (1998).
- Gottschling, D. E., Aparicio, O. M., Billington, B. L. & Zakian, V. A. Position effect at *S. cerevisiae* telomeres: reversible repression of Pol II transcription. *Cell* **63**, 751–762 (1990).
- Gotta, M. *et al.* The clustering of telomeres and colocalization with Rap1, Sir3, and Sir4 proteins in wild-type *Saccharomyces cerevisiae*. *J. Cell Biol.* **134**, 1349–1363 (1996).
- Laroche, T. *et al.* Mutation of yeast Ku genes disrupts the subnuclear organization of telomeres. *Curr. Biol.* **8**, 653–656 (1998).
- Feldmann, H. *et al.* HDF2, the second subunit of the Ku homologue from *Saccharomyces cerevisiae*. *J. Biol. Chem.* **271**, 27765–27769 (1996).
- Gravel, S., Larrière, M., Labrecque, P. & Wellinger, R. J. Yeast Ku as a regulator of chromosomal DNA end structure. *Science* **280**, 741–744 (1998).
- Milne, G. T., Jin, S., Shannon, K. B. & Weaver, D. T. Mutations in two Ku homologs define a DNA end-joining repair pathway in *Saccharomyces cerevisiae*. *Mol. Cell Biol.* **16**, 4189–4198 (1996).
- Boulton, S. J. & Jackson, S. P. *Saccharomyces cerevisiae* Ku70 potentiates illegitimate DNA double-strand break repair and serves as a barrier to error-prone DNA repair pathways. *EMBO J.* **15**, 5093–5103 (1996).
- Moore, J. K. & Haber, J. E. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in *Saccharomyces cerevisiae*. *Mol. Cell Biol.* **16**, 2164–2173 (1996).
- Barnes, G. & Rio, D. DNA double-strand-break sensitivity, DNA replication, and cell cycle arrest phenotypes of Ku-deficient *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **94**, 867–872 (1997).
- Porter, S. E., Greenwell, P. W., Ritchie, K. B. & Petes, T. D. The DNA-binding protein Hdf1p (a putative Ku homologue) is required for maintaining normal telomere length in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **24**, 582–585 (1996).
- Boulton, S. J. & Jackson, S. P. Components of the Ku-dependent non-homologous end-joining pathway are involved in telomeric length maintenance and telomeric silencing. *EMBO J.* **17**, 1819–1828 (1998).
- Nugent, C. I. *et al.* Telomere maintenance is dependent on activities required for end repair of double-strand breaks. *Curr. Biol.* **8**, 657–660 (1998).
- Polotnianka, R. M., Li, J. & Lustig, A. J. The yeast Ku heterodimer is essential for protection of the telomere against nucleolytic and recombinational activities. *Curr. Biol.* **8**, 831–834 (1998).
- Cordes, V. C., Reidenbach, S., Rackwitz, H. R. & Franke, W. W. Identification of protein p270/Tpr as a constitutive component of the nuclear pore complex-attached intranuclear filaments. *J. Cell Biol.* **136**, 515–529 (1997).
- Bangs, P. *et al.* Functional analysis of Tpr: identification of nuclear pore complex association and nuclear localization domains and a role in mRNA export. *J. Cell Biol.* **143**, 1801–1812 (1998).
- Kolling, R., Nguyen, T., Chen, E. Y. & Botstein, D. A new yeast gene with a myosin-like heptad repeat structure. *Mol. Gen. Genet.* **237**, 359–369 (1993).
- Strambio-de-Castillia, C., Blobel, G. & Rout, M. P. Proteins connecting the nuclear pore complex with the nuclear interior. *J. Cell Biol.* **144**, 839–855 (1999).
- Teixeira, M. T., Fabre, E. & Dujon, B. Self catalysed cleavage of the yeast nucleoporin Nup145p precursor. *J. Biol. Chem.* **274**, 32439–32444 (1999).
- Martin, S. G., Laroche, T., Suka, N., Grunstein, M. & Gasser, S. M. Relocalization of telomeric Ku and SIR proteins in response to DNA strand breaks in yeast. *Cell* **97**, 621–633 (1999).

- Shore, D. Telomere length regulation: getting the measure of chromosome ends. *Biol. Chem.* **378**, 591–597 (1997).
- Schimmang, T., Tollervey, D., Kem, H., Frank, R. & Hurt, E. C. A yeast nucleolar protein related to mammalian fibrillarin is associated with small nucleolar RNA and is essential for viability. *EMBO J.* **8**, 4015–4025 (1989).
- Lee, S. E. *et al.* *Saccharomyces* Ku70, mre11/rad50 and RPA proteins regulate adaptation to G2/M arrest after DNA damage. *Cell* **94**, 399–409 (1998).
- Trelles-Sticken, E., Loidl, J. & Scherthan, H. Bouquet formation in budding yeast: initiation of recombination is not required for meiotic telomere clustering. *J. Cell Sci.* **112**, 651–658 (1999).
- Blobel, G. Gene gating: a hypothesis. *Proc. Natl Acad. Sci. USA* **82**, 8527–8529 (1985).
- Mages, G. J., Feldmann, H. M. & Winnacker, E. L. Involvement of the *Saccharomyces cerevisiae* HDF1 gene in DNA double-strand break repair and recombination. *J. Biol. Chem.* **271**, 7910–7915 (1996).
- Kilmartin, J. V. & Adams, A. E. Structural rearrangements of tubulin and actin during the cell cycle of the yeast *Saccharomyces*. *J. Cell Biol.* **98**, 922–933 (1984).
- Aris, J. P. & Blobel, G. Yeast nuclear envelope proteins cross react with an antibody against mammalian pore complex proteins. *J. Cell Biol.* **108**, 2059–2067 (1989).
- Olivo, J.-C. Automatic detection of spots in biological images by a wavelet-based selective filtering technique. *Proc. IEEE Int. Conference on Image Processing ICIP 1996*, 311–314 (1996).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Correspondence and requests for materials should be addressed to U.N.

Three-dimensional structure of the ion-coupled transport protein NhaA

Karen A. Williams

Department of Structural Biology, Max Planck Institute of Biophysics,
Heinrich Hoffmann Str. 7, D-60528 Frankfurt/Main, Germany

Ion-coupled membrane-transport proteins, or secondary transporters, comprise a diverse and abundant group of membrane proteins that are found in all organisms. These proteins facilitate solute accumulation and toxin removal against concentration gradients using energy supplied by ion gradients across membranes. NhaA is a Na^+/H^+ antiporter of relative molecular mass 42,000, which is found in the inner membrane of *Escherichia coli*, and which has been cloned and characterized^{1,2}. NhaA uses the H^+ electrochemical gradient to expel Na^+ from the cytoplasm, and functions primarily in the adaptation to high salinity at alkaline pH^{1,2}. Most secondary transporters, including NhaA³, are predicted to have 12 transmembrane helices. Here we report the structure of NhaA, at 7 Å resolution in the membrane plane and at 14 Å vertical resolution, determined from two-dimensional crystals⁴ using electron cryo-microscopy. The three-dimensional map of NhaA reveals 12 tilted, bilayer-spanning helices. A roughly linear arrangement of six helices is adjacent to a compact bundle of six helices, with the density for one helix in the bundle not continuous through the membrane. The molecular organization of NhaA represents a new membrane-protein structural motif and offers the first insights into the architecture of an ion-coupled transport protein.

Secondary transporters have central roles in human health and disease (for example, the serotonin transporter in nervous system

Table 1 Crystallographic data summary

Plane group	$P2_12_1$
Unit-cell dimensions	$a = 47.5 \text{ Å}, b = 181.4 \text{ Å}, \gamma = 90^\circ$
No. of images*	44
Total no. of observations†	9,709
No. of unique observations	1,954
Overall weighted phase error to 7.0 Å	18.7°

* Distribution of tilt angles: 5 (0°); 13 (20°); 17 (30°); 8 (45°).

† Including reflections with signal-to-noise ratios > 1 (IQ 1–8). Three representative lattice lines are available as Supplementary Information.

function^{5,6}; the Na⁺/glucose co-transporter SGLT1 in glucose/galactose malabsorption⁷; and antibiotic resistance⁸). Insights into the molecular mechanism of ion-coupled transport have come from studies of many secondary transporters, most notably the extensive site-directed biophysical studies on lactose permease⁹, an *E. coli* H⁺/galactoside symporter. A complete understanding of ion-coupled transport nonetheless requires the framework provided by a knowledge of the 3D structure, which has been hindered by the difficulty in growing well-ordered crystals of these proteins. Highly ordered 2D crystals of NhaA, a representative secondary transporter, have recently been obtained⁴ and are ideal for structure determination by electron crystallography. A projection map at 4 Å resolution showed that each unit cell contains four molecules of NhaA organized as dimers in alternately up-and-down orientations, and that NhaA exhibits a highly asymmetrical structure⁴. To resolve the features present in the projection map, a series of images of crystals tilted at angles from 0° to 45° were collected, and the amplitudes and phases determined from these images were used to calculate a 3D map of NhaA. The map was calculated with data truncated at ~7 Å resolution in the membrane plane and at ~14 Å perpendicular to the membrane (Table 1; and Supplementary Information).

Twelve rod-shaped density features indicative of transmembrane (TM) helices can be discerned in NhaA (Fig. 1). Twelve density peaks can also be seen in each of three slices taken at different heights through the membrane (Fig. 2). It is not clear whether the view shown in Fig. 1 corresponds to the periplasmic or cytoplasmic surface of the protein. All of the helices in NhaA are well resolved, with the exception of a group of three helices that includes one helix for which density is not continuous (Fig. 3c, discussed below).

The boundary of the NhaA monomer predicted from the projection map⁴ was confirmed in the 3D map on the basis of the large distance between the two monomers in a dimer on one side of the membrane (Fig. 2a). Close contacts between the two monomers occur principally between two pairs of helices on the other side of the membrane (Fig. 2c). The boundary between adjacent dimers was previously established by the comparison of projection maps

from two different crystal forms⁴. The organization of helices on one side of the membrane differs markedly from that on the opposite surface. There is a compact arrangement of helices in a roughly triangular overall shape on one side (Fig. 2a), whereas on the opposite surface the helices are loosely packed into a more irregular arrangement (Fig. 2c).

Although NhaA crystallizes as a dimer, it is not clear whether the monomer or dimer is functionally active. Nonetheless, the large distance between the two monomers on one side of the membrane (Fig. 2a) suggests that the substrate translocation pathway does not form at the dimer interface. In contrast to the intradimer contacts that occur on only one side of the membrane, helix-helix contacts between adjacent dimers extend through the membrane (shown by the white arrows in Fig. 4) and are probably important in contributing to the high degree of order exhibited by these crystals.

To examine NhaA in more detail, it can be divided into two principal regions: a layer of six helices mediating the dimer interface (Fig. 3a); and an adjacent bundle of six helices (Fig. 3b, c). The average thickness of NhaA is ~45 Å (contoured at 1σ; standard deviation), which is close to the thickness of most biological membranes. The six helices mediating dimer contacts seem to be on average shorter (~32 Å) than the helices in the adjacent bundle (~40 Å). Helices in NhaA are packed in a predominantly right-handed manner and are tilted ~10–27° relative to the membrane normal, with the average tilt being ~20°. The substantial tilt of most helices makes it clear why elongated regions of density rather than discrete peaks predominate in the projection density map.

The six helices that mediate dimer contacts are tilted in the same direction towards the left side of the panel (Fig. 3a). One of the helices has a striking 'L' shape, which may represent a TM helix connected to a short helix oriented parallel to the membrane surface, or perhaps a compact loop. A feature on the upper surface showing spherical density suggests the presence of another ordered loop.

The helices in the bundle are most clearly seen by first viewing the layer of three long helices (Fig. 3b), which are tilted in the direction

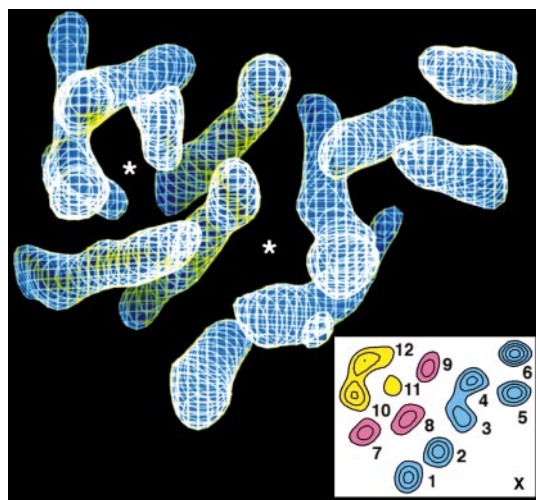


Figure 1 The 3D map of NhaA at 7 Å resolution viewed normal to the membrane plane. Twelve rod-shaped density features corresponding to transmembrane helices can be seen, with the density for one of the helices not continuous through the bilayer. Two possibilities for the location of ion-translocation pathways are shown by asterisks (see text for discussion). Inset, horizontal section taken from the centre of the 3D map with the protein in the same orientation and helices arbitrarily numbered from 1 to 12 (no correspondence to helix assignments in the protein). The side views of the helices shown in Fig. 3 are from the perspective of 'X' in the inset. The blue helices are displayed in Fig. 3a, magenta helices in Fig. 3b and yellow helices in Fig. 3c. The map was contoured at 3σ above the mean.

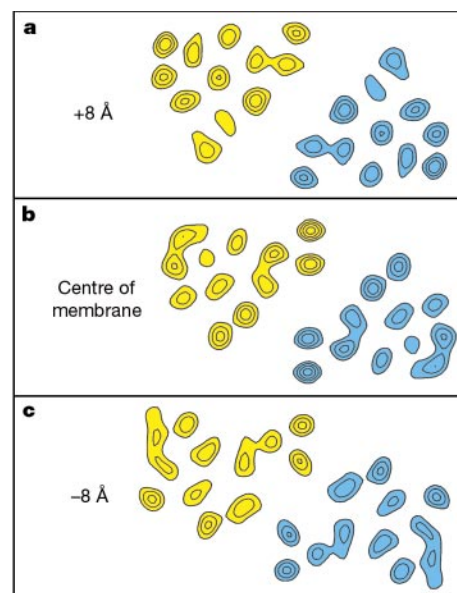


Figure 2 Horizontal slices through the NhaA dimer. Monomers are shown in yellow or blue. **a**, 8 Å above the centre of the membrane; **b**, at the centre of the membrane; and **c**, 8 Å below the centre of the membrane. In each slice, 12 peaks are seen for a monomer of NhaA.

opposite to that of the first layer of six helices, and then viewing the group of three incompletely resolved helices (Fig. 3c). The comparatively lower resolution seen in Fig. 3c makes it difficult to interpret the density features definitively. The limited resolution may arise from conformational heterogeneity, and, indeed, this region of NhaA lacks the close interdimer and intradimer contacts that may stabilize the remainder of the molecule. Nonetheless, it seems likely that there are two long helices and a third helix which consists of two rod-shaped density features (Figs 1, 3c). Although this is most simply interpreted as a continuous TM helix, it is possible that there are, in fact, two helical regions connected by a more extended or disordered structure.

The molecular organization of NhaA represents a new structural motif. It bears no resemblance to the 12 TM-spanning helices of subunit I of cytochrome *c* oxidase, which displays apparent three-fold symmetry with helices organized into three 4-helix bundles¹⁰. There are several interesting parallels between the 3D structure of NhaA and that of the P-type ATPases that couple ATP hydrolysis with ion transport. The H⁺-ATPase¹¹ and Ca²⁺-ATPase¹² both have 10 TM helices which are organized into a right-handed 4-helix bundle adjacent to a layer of four helices which are in turn covered by a further layer of two helices. NhaA can be considered as having a right-handed bundle of helices adjacent to a layer of four helices, with a further layer of two helices. But in NhaA the bundle is comprised of six rather than four helices, the tilts of the layer of four helices are in the opposite direction to the helices in the P-type ATPases, and the most distal layer of two helices is disposed to one

side of the molecule rather than being in the centre. NhaA bears little resemblance to two other ion pumps, bacteriorhodopsin^{13,14} and halorhodopsin¹⁵, both of which occur as bundles of seven helices, with the chromophore retinal in the centre. The TM helix organization of NhaA also differs from that of channel proteins that facilitate bi-directional substrate flow across the membrane and frequently exhibit a high degree of symmetry (for example, the tetrameric KscA K⁺ channel¹⁶ and the pentameric MscL mechanosensitive channel¹⁷). Perhaps the closest comparison between NhaA and a channel protein can be made to the AQP1 water channel, which occurs as a right-handed 6-helix barrel, with additional density in the centre corresponding to two loops inserted into the barrel^{18–20}. NhaA shows a more compact right-handed bundle of six helices, with the helix comprised of two rod-shaped density features crossing the centre of the bundle.

The 3D map of NhaA was determined from 2D crystals of NhaA obtained at low pH, where NhaA is in a conformation associated with the inactive state²¹. It may be possible to examine the conformational changes associated with transport by exposing the crystals to a shift in pH or Na⁺ concentration and then trapping the conformational change by rapid freezing (for example, similar to the trapping of the acetylcholine-induced conformational changes achieved with the acetylcholine receptor²²). This would be valuable in establishing whether the conformational changes are similar to the localized changes that occur during the bacteriorhodopsin photocycle²³, or the helix translations and rotations predicted to occur upon the opening of the K⁺ channel based on site-directed spin-labelling methods²⁴. Assuming that the conformation of the protein in the inactive state is not markedly different from that of the active protein, the map suggests two possibilities for the location of ion-translocation pathways: within the bundle of six helices; and between the bundle of six helices and the adjacent layer of six helices (marked by asterisks in Fig. 1).

NhaA is a member of an extremely large group of secondary transporters which differ in size and topology, substrates, coupling ions and the directionality of transport (that is, both symport and antiport). Although many members of this group are thought to have arisen from an internal duplication of six TM helices, many others, such as NhaA, show no evidence of this. It will be intriguing to learn whether the structure adopted by NhaA is also adopted in full, or in part, by other members of this group of transport

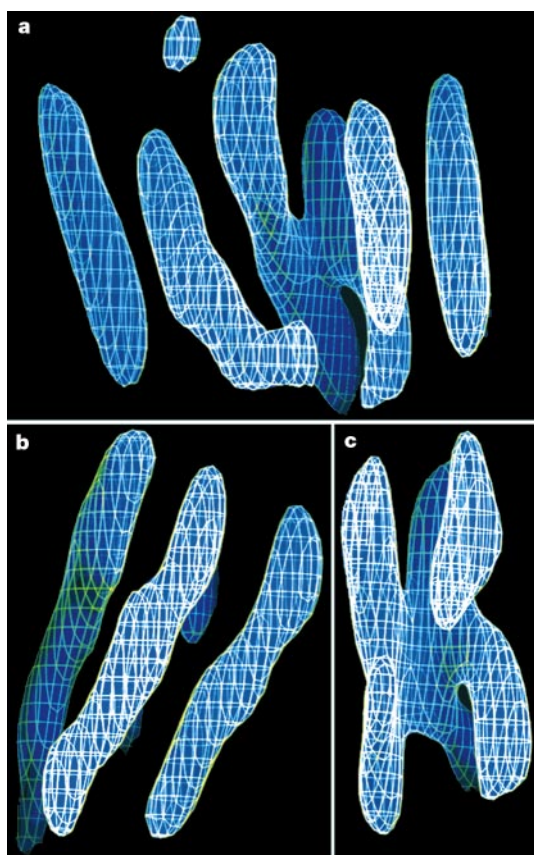


Figure 3 Vertical sections through the NhaA 3D map. The molecule is rotated 90° to reveal the helices perpendicular to the membrane from the perspective of the 'X' in the horizontal section taken from the centre of the membrane in Fig. 1 inset. **a**, The six helices that mediate intradimer and interdimer contacts (blue in Fig. 1 inset); **b**, the three long tilted helices (magenta in Fig. 1 inset) that are revealed when the first layer of six helices is removed; and **c**, the group of three incompletely resolved helices (yellow in Fig. 1 inset) that are seen when the layer of three helices in **b** is removed.

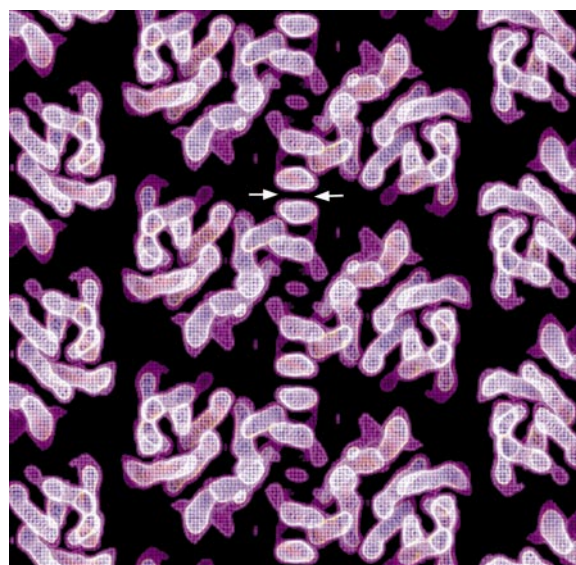


Figure 4 Overview of the interactions between adjacent dimers of NhaA. The map is contoured in magenta at 1.5 σ over white contours at 3 σ . The white arrows show close interdimer contacts between helices of adjacent dimers.

proteins. Higher resolution structural information is required to shed light on the molecular mechanism of ion-coupled transport, specifically the ion specificity and activity of NhaA. It appears feasible that this will be obtained given that the crystals diffract to near atomic resolution⁴. □

Methods

Two-dimensional crystals of NhaA were obtained as described⁴ and, in addition, the NhaA concentration was generally lowered to 0.5 mg ml⁻¹ and the *E. coli* polar lipids (Avanti) were solubilized in 0.15% decyl maltoside (Sigma) in the starting crystallization mixture. Frozen hydrated specimens were prepared by the back-injection of the crystals into a tannin or trehalose solution²⁵. Images were recorded using a JEOL 3000 SFF electron microscope equipped with a field-emission gun and liquid-helium-cooled top-entry stage using 0°, 20°, 30° or 45° tilt holders, with an accelerating voltage of 300 kV at a nominal specimen temperature of 4 K and a magnification of ×53,300 or ×70,000. Images were recorded either in flood-beam mode with a 1-s exposure time at an estimated electron dose of 20–30 electrons Å⁻², or using a spot-scan procedure with an exposure time of 40 ms per spot and a similar total dose. Images were recorded on Kodak SO-163 electron emulsion film and developed for 12 min in full-strength Kodak D19 developer. Images were evaluated by optical diffraction, and well-ordered areas of 6,000 × 6,000 pixels corresponding to ~4 cm² on the negatives were digitized using a 7-μm pixel size on a Zeiss SCAI scanner. Images were processed using MRC image processing programs to correct for lattice distortions and the effects of the contrast transfer function^{26,27}. Initial estimates of tilt angles were calculated from lattice parameters²⁸ and the images merged in the plane group P2₂1₂1. Phase origins, tilt geometry and contrast transfer function were refined in two rounds, first to 15 Å and then to 7 Å against the model. The point-spread function²⁹ indicated a nominal maximum resolution of ~7 Å in the membrane plane and ~14 Å perpendicular to it. Image amplitudes were scaled with an average temperature factor *B* = -350 to compensate for the resolution-dependent degradation of amplitudes using SCALIMAMP3D (ref. 15). The 3D density map was calculated with CCP4 and the map displayed with the graphics program O (ref. 30).

Received 1 September; accepted 15 November 1999.

1. Padan, E. *et al.* The molecular mechanism of regulation of the NhaA Na⁺/H⁺ antiporter of *Escherichia coli*, a key transporter in the adaptation to Na⁺ and H⁺. *Novartis Found. Symp.* **221**, 183–196 (1999).
2. Padan, E. & Schuldiner, S. in *Handbook of Biological Physics* (eds Konings, W. N., Kaback, H. R. & Lolkema, J. S.) 501–531 (Elsevier, Oxford, 1996).
3. Rothman, A., Padan, E. & Schuldiner, S. Topological analysis of NhaA, a Na⁺/H⁺ antiporter from *Escherichia coli*. *J. Biol. Chem.* **271**, 32288–32292 (1996).
4. Williams, K. A. *et al.* Projection structure of NhaA, a secondary transporter from *Escherichia coli*, at 4.0 Å resolution. *EMBO J.* **18**, 3558–3563 (1999).
5. Blakely, R. D., De Felice, L. J. & Hartzell, H. C. Molecular physiology of norepinephrine and serotonin transporters. *J. Exp. Biol.* **196**, 263–281 (1994).
6. Schloss, P. & Williams, D. C. The serotonin transporter: a primary target for antidepressant drugs. *J. Psychopharmacol.* **12**, 115–121 (1998).
7. Wright, E. M. Glucose galactose malabsorption. *Am. J. Physiol.* **275**, 879–882 (1998).
8. Saier, M. H. Jr *et al.* Evolutionary origins of multidrug and drug-specific efflux pumps in bacteria. *FASEB J.* **12**, 265–274 (1998).
9. Kaback, H. R. & Wu, J. From membrane to molecule to the third amino acid from the left with a membrane transport protein. *Q. Rev. Biophys.* **30**, 333–364 (1997).

10. Iwata, S., Ostermeier, C., Ludwig, B. & Michel, H. Structure at 2.8 Å resolution of cytochrome *c* oxidase from *Paracoccus denitrificans*. *Nature* **376**, 660–669 (1995).
11. Auer, M., Scarborough, G. A. & Kühlbrandt, W. Three-dimensional map of the plasma membrane H⁺-ATPase in the open conformation. *Nature* **392**, 840–843 (1998).
12. Zhang, P. *et al.* Structure of the calcium pump from sarcoplasmic reticulum at 8 Å resolution. *Nature* **392**, 835–839 (1998).
13. Henderson, R. *et al.* Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* **213**, 899–929 (1990).
14. Grigorieff, N., Ceska, T. A., Downing, K. H., Baldwin, J. M. & Henderson, R. Electron-crystallographic refinement of the structure of bacteriorhodopsin. *J. Mol. Biol.* **259**, 393–421 (1996).
15. Havelka, W. A., Henderson, R. & Oesterhelt, D. Three-dimensional structure of halorhodopsin at 7 Å resolution. *J. Mol. Biol.* **247**, 726–738 (1995).
16. Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K⁺ conduction and selectivity. *Science* **280**, 69–77 (1998).
17. Chang, G., Spencer, R. H., Lee, A. T., Barclay, M. T. & Rees, D. C. Structure of the MscL homolog from *Mycobacterium tuberculosis*: a gated mechanosensitive ion channel. *Science* **282**, 2220–2226 (1998).
18. Walz, T. *et al.* The three-dimensional structure of aquaporin-1. *Nature* **387**, 624–627 (1997).
19. Cheng, A. *et al.* Three-dimensional organization of a human water channel. *Nature* **387**, 627–630 (1997).
20. Li, H., Lee, S. & Jap, B. K. Molecular design of aquaporin-1 water channel as revealed by electron crystallography. *Nature Struct. Biol.* **4**, 263–265 (1997).
21. Rothman, A., Gerchman, Y., Padan, E. & Schuldiner, S. Probing the conformation of NhaA, a Na⁺/H⁺ antiporter from *Escherichia coli*, with trypsin. *Biochemistry* **36**, 14572–14576 (1997).
22. Unwin, N. Acetylcholine receptor in the open state. *Nature* **373**, 37–43 (1995).
23. Subramaniam, S. *et al.* Protein conformational changes in the bacteriorhodopsin photocycle. *J. Mol. Biol.* **287**, 145–161 (1999).
24. Peruse, E., Courts, D. M. & Cull, L. G. Structural rearrangements underlying K⁺ channel activation gating. *Science* **285**, 73–78 (1999).
25. Wang, D. N. & Kühlbrandt, W. High-resolution electron crystallography of light-harvesting chlorophyll *a/b*-protein complex in three different media. *J. Mol. Biol.* **217**, 691–699 (1991).
26. Henderson, R. *et al.* Structure of purple membrane from *Halobacterium halobium*: recording, measurement and evaluation of electron micrographs at 3.5 Å resolution. *Ultramicroscopy* **19**, 899–929 (1986).
27. Crowther, R. A., Henderson, R. & Smith, J. M. MRC image processing programs. *J. Struct. Biol.* **116**, 9–16 (1996).
28. Shaw, P. J. & Hills, G. J. Tilted specimen in the electron microscope: a simple specimen holder and the calculation of tilt angles from crystalline specimens. *Micron* **12**, 279–282 (1981).
29. Unger, V. M. & Schertler, G. F. X. Low resolution structure of bovine rhodopsin determined by electron cryo-microscopy. *Biophys. J.* **68**, 1776–1786 (1995).
30. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps. *Acta Crystallogr. A* **47**, 110–119 (1991).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial offices of Nature.

Acknowledgements

I thank W. Kühlbrandt for his support and encouragement throughout this project, D. Mills for assistance with the JEOL 3000 SFF, V. Unger for advice on image processing, U. Geldmacher-Kaufer for technical assistance, and S. Schuldiner, E. Padan and A. Rothman for their assistance with NhaA biochemistry during the early stages of this work. This work was supported by the Deutsche Forschungsgesellschaft.

Correspondence and requests for materials should be addressed to K.A.W. (e-mail: williams@biophys.mpg.de).